Knowledge Graph Creation for Volunteered Geographic Information

Kumulative Dissertation
zur
Erlangung des Doktorgrades (Dr. rer. nat.)
der
Mathematisch-Naturwissenschaftlichen Fakultät
der
Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von
Alishiba Florian Dsouza
aus
Agashi, India

Bonn, 2024

Angefertigt mit Genehmigung
der Mathematisch-Naturwissenschaftlichen Fakultät
der Rheinischen Friedrich-Wilhelms-Universität Bonn

Gutachterin/Betreuerin: Prof. Dr. Elena Demidova
Gutachterin: Prof. Dr. Anna Fensel

Tag der Promotion: 05.07.2024
Erscheinungsjahr: 2024

ABSTRACT

Community-created geographic data sources, such as OpenStreetMap, are semi-structured sources of geographic information. These sources rely on voluntary contributions from individuals worldwide, resulting in huge amounts of geographic data. However, this volunteer nature poses challenges due to the varying interests and expertise of the volunteers. The entity and schema representations within OpenStreetMap are sparse and heterogeneous, making it challenging to manage and access the data for downstream applications. Meanwhile, semantic data sources such as knowledge graphs offer more structured data, following ontologies that provide uniform representations and semantic relationships. Aligning community-generated geographical resources with knowledge graphs allows us to enrich the former with semantic data and provide the knowledge graphs with precise geographical information.

However, integrating volunteered geographic information sources and knowledge graphs is difficult due to challenges such as less annotated data, poor data quality, and representational differences between entities and schema. To alleviate these obstacles, in this thesis, we propose solutions for the alignment of entities and schemas and creating a comprehensive geographic knowledge graph. Initially, we present NCA, a neural model to align OpenStreetMap schema elements, commonly referred to as tags, with knowledge graph classes by utilizing a novel shared latent space and contrastive learning. Then, by utilizing the knowledge gained from NCA, we present IGEA, an iterative approach to align schema elements and entities between OSM and knowledge graphs. IGEA leverages a cross-attention mechanism for the alignment. By utilizing entity descriptions from multiple sources, IGEA finds better alignments than the state-of-the-art approaches. Finally, we present WorldKG, a novel geographic knowledge graph containing OpenStreetMap data in semantic format. WorldKG knowledge graph adheres to the novel WorldKG ontology created by representing the tags of OSM in superclass subclass relations.

By addressing the challenges of data integration and implementing a structured ontology, WorldKG serves as a valuable resource for downstream applications, providing a platform for accessing and leveraging geographic data in a structured and comprehensive manner. In addition to the immediate benefits for downstream applications, the methods and knowledge graph developed in this thesis will benefit further developments in the domain of geographic information on the web.

Keywords: *Volunteered Geographic Information, Geographic Schema Alignment, Geographic Entity Alignment, Geographic Knowledge Graphs*

Acknowledgements

I extend my gratitude to Prof. Dr. Elena Demidova for her invaluable supervision and support throughout my Ph.D. journey. I am thankful to Prof. Dr. Anna Fensel for agreeing to serve as the second reviewer for my thesis. Additionally, I want to thank Prof. Dr. Thomas Schultz and Prof. Dr. Jochen Dingfelder for being part of the doctoral committee.

I would like to thank all my colleagues from DSIS group for their support. Finally, I want to thank my family and my amazing husband Rakesh for their support, love, and care.

The works presented in the chapters of this thesis were partially funded by the DFG, German Research Foundation ("WorldKG", 424985896), the Federal Ministry for Economic Affairs and Climate Action (BMWK), Germany ("ATTENTION!", 01MJ22012C), and DAAD/BMBF, Germany ("KOALA", 57600865).

List of Publications

In this thesis, I have contributed towards the completion of geographic knowledge on the web. The contributions presented in this thesis have been published at the following conferences:

- **Alishiba Dsouza**, Nicolas Tempelmeier, and Elena Demidova. "Towards Neural Schema Alignment for OpenStreetMap and Knowledge Graphs". In: Proceeding of the 20th International Semantic Web Conference, ISWC 2021. Springer, 2021, pp. 56–73, DOI: 10.1007/978-3-030-88361-4_4
- **Alishiba Dsouza**, Nicolas Tempelmeier, Ran Yu, Simon Gottschalk, Elena Demidova. "WorldKG: A World-Scale Geographic Knowledge Graph". In: Proceeding of the 30th ACM International Conference on Information and Knowledge Management, CIKM 2021. ACM, 2021, pp. 4475–4484. DOI: 10.1145/3459637.3482023.
- **Alishiba Dsouza**, Ran Yu, Moritz Windoffer, Elena Demidova. "Iterative Geographic Entity Alignment with Cross Attention". In: Proceedings of the 22nd International Semantic Web Conference, ISWC 2023. Springer, 2023, pp. 216–233. DOI: 10.1007/978-3-031-47240-4_12. *Best Student Paper Award*

In addition to the works mentioned above, I also contributed to the following publications that are not part of this thesis.

- Genivika Mann, **Alishiba Dsouza**, Ran Yu, Elena Demidova. "Spatial Link Prediction with Spatial and Semantic Embeddings". In: Proceedings of the 22nd International Semantic Web Conference, ISWC 2023. Springer, 2023, pp. 179–196. DOI: 10.1007/978-3-031-47240-4_10. *Best Paper Award*.
- **Alishiba Dsouza**, Nicolas Tempelmeier, Simon Gottschalk, Ran Yu, and Elena Demidova. "WorldKG: World-Scale Completion of Geographic Information". In: Volunteered Geographic Information. Springer, 2023, pp. 3–19. ISBN: 9783031353741. DOI: 10.1007/978-3-031-35374-1_1.
- Hamed Aboutorab, Ran Yu, **Alishiba Dsouza**, Morteza Saberi, Omar Khadeer Hussain "News Recommendation System for Environmental Risk Management". In: Proceedings of the Second International Workshop on Linked Data-driven Resilience Research 2023 co-located with Extended Semantic Web Conference 2023. Vol. 3401. CEUR Workshop Proceedings. 2023.
- **Alishiba Dsouza**, Moritz Schott, and Sven Lautenbach. "Comparative Integration Potential Analyses of OSM and Wikidata–The Case Study of Railway Stations". In: Proceedings of the Academic Track at State of the Map, 2022.
- Maximilian Hartmann, Moritz Schott, **Alishiba Dsouza**, Yannick Metz, Michele Volpi, and Ross S. Purves. Text and Image Analysis Workflow using Citizen Science Data to Extract Relevant Social Media Records: Combining

Red Kite Observations from Flickr, eBird and iNaturalist". In: Ecological Informatics 71, 2022, p. 101782. DOI: 10.1016/J.ECOINF.2022.101782.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

Geographic information has gained a lot of attention in recent years due to applications such as POI recommendation, navigation, and routing applications. The use of geographic information in the form of maps dates back centuries, but with the rise in internet usage, the number of applications requiring geographic information has also grown. Over the last few decades, many sources of geographic data have emerged. In 2007, Goodchild [Goo07] introduced the term volunteered geographic information (VGI). In VGI, the data is collected and maintained by volunteers. In 2006, OpenStreetMap (OSM) emerged as the source of free geographic data and is currently one of the biggest sources of volunteered geographic information, with over 11 million registered users and over 1.4 million editors[1]. OSM has vast geographic objects in various forms such as points, lines, and polygons, and the objects are described using key=value pairs called ´tags'.

Although the OSM data is exceedingly large, data is sparse and heterogeneous due to the voluntary nature of the data collection. For example, in March 2024, in the country of Germany, there were over 400 million nodes (points) created out of which only about 4.5% of nodes had at least one tag with an average of 4 tags per node[2]. Objects of certain types, such as cities, and famous railway stations, are described in more detail, whereas lesser utilized types such as small villages are rarely described. Furthermore, the data statistics vary depending on the regions and countries. For example, contrary to Germany, in the country of India, there were over 200 million nodes (points) created, out of which only about 1.3% of nodes had at least one tag with an average of 2 tags per node[3]. To create OSM objects, volunteers are provided with a set of guidelines. These guidelines, however, do not conform to a fixed ontology, making the OSM schema ever-growing and highly heterogeneous. OSM data in its original form cannot be made directly available for semantic applications such as geographic question answering and information retrieval.

Knowledge graphs, since their inception in 2012, have been popular as a source of semantic information. Unlike OSM, knowledge graphs follow a strict ontology and can be made directly accessible to semantic applications. Although general-purpose knowledge graphs such as Wikidata, and DBpedia contain structured information about geographic data, their coverage is not extensive. Figure 1.1 shows the current geographic entities present in OSM (over 400 million[4]) and the Wikidata knowledge graph (over 0.8 million[5]) for the country of Germany. As

[1] https://osmstats.neis-one.org/
[2] https://taginfo.geofabrik.de/germany/reports/database_statistics
[3] https://taginfo.geofabrik.de/india/reports/database_statistics
[4] https://download.geofabrik.de/europe/germany.html
[5] https://query.wikidata.org/

seen in the figure, the geographic entities in Wikidata are fewer than in OSM. Even domain-specific knowledge graphs such as LinkedGeoData [ALH09] and Yago2Geo [KMK19] lack in terms of size and types. Integrating OpenStreetMap and knowledge graphs can benefit geographic data users with precise geographic information from OSM and structured and semantic geographic information from knowledge graphs.

The integration of OSM and knowledge graphs is a very challenging task. Following, in Table 1.1 we show an example of the entity Mount Everest in Wikidata knowledge graph and OSM to understand the differences between these sources.

TABLE 1.1: Representation of Mount Everest in OpenStreetMap and Wikidata

Key	Value
id	164979149
name	Mount Everest
natural	peak
way	POINT (27.988 08 86.92514)
ele	8848.86

A) OpenStreetMap tags

Subject	Predicate	Object
`Q513`	*label*	Mount Everest
`Q513`	*instance of*	`Q8502` (*mountain*)
`Q513`	*location*	`Q5451` (*Himalayas*)
`Q513`	*coordinate location*	27°59'17.6500"N, 86°55'30.0652"E
`Q513`	*elevation above sea level*	*8,844.43 metre*

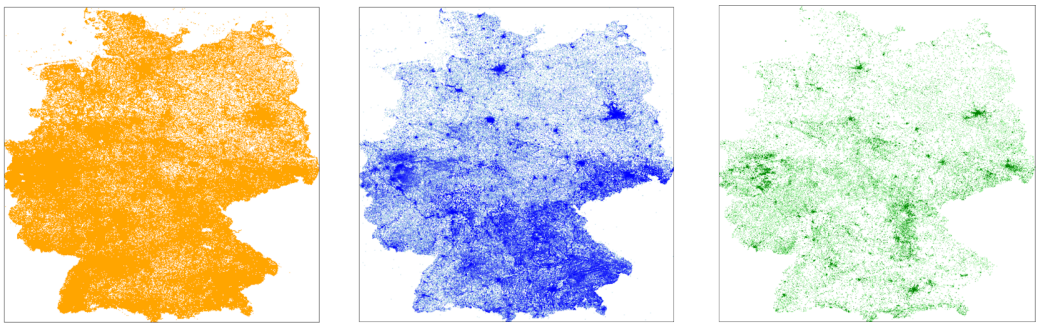
B) Wikidata triples. `Q513` represents Mount Everest

As seen in Table 1.1, the Wikidata knowledge graph has an *instance of* property that describes the type of entity. In OSM, unless we have knowledge of the schema, it is difficult to know what is the type of given entity. Moreover, there exists a many-to-many relation between OSM and KG classes. For example, OSM tags *place=city* can be mapped to Wikidata classes such as *City, Big City, and Capital City*. Furthermore, there exists only a fraction of already linked schema elements between OSM and knowledge graphs which is not sufficient to train supervised models. The OSM schema is flat and does not have any hierarchical relations between its schema elements. Such challenges including representational differences, ambiguities, and lack of data hinder the seamless schema alignment. The alignment at the entity level is challenging, since the values of entities are not similar. In the given example, the values of elevation and geographic coordinates vary between the two sources, hindering the application of similarity-based approaches for alignment.

This thesis proposes ways to integrate OSM and knowledge graphs at schema and entity levels and to lift OSM into a semantic representation. We first tackle the problem of schema alignment by aligning tags of OSM to classes of knowledge graphs. Then, we align the entities of these two sources and finally, we convert the OSM data into a semantically structured data source.

1.1 Research Questions

As explained earlier, the OSM schema is heterogeneous. Moreover, the semantics of the tags are unclear. There is no clear distinction between tags describing the type of the object and tags describing other properties. Aligning the schema elements becomes inherently difficult due to the unclear semantics and flatness of the OSM schema, which brings us to our first research question.



A) OSM B) Wikidata C) OSM-Wikidata

FIGURE 1.1: Geographic entities in OSM, Wikidata, and linked entities between Wikidata and OSM for the country of Germany

RQ1. How to create a neural model to link OpenStreetMap and Knowledge graphs at the schema level?

Linking OSM and knowledge graphs is a challenging task. Currently, there are very few linked schema elements between OSM and knowledge graphs, which limits the use of machine learning models relying on massive data for training. Furthermore, the linked entities between these sources are very sparse, as seen in Figure 1.1. Using simple string similarity measures to link schema elements is not enough, as OSM and knowledge graphs follow different naming schemes. For example, the OSM tag *natural=peak* describes an object of type mountain. In Wikidata knowledge graph, entities of type mountain are described with the type *Q8502 (Mountain)*. Using string similarity measures, *natural=peak* and *Mountain*, do not yield accurate results. There exists a need to create a model that incorporates the semantics of the tags along with the rich information provided by OSM and knowledge graphs to get the alignments between OSM tags and KG classes.

The tasks of schema and entity alignment are interlinked. Having more linked entities can enhance the task of schema alignment, and having more schema elements linked can help in better entity alignment. This trade-off brings us to our second research question.

2. How to identify identity links between OSM and knowledge graphs using an iterative neural model?

As seen in Figure 1.1, there is a huge linking potential between OSM and general-purpose knowledge graphs. Due to the schema mismatch between OSM and knowledge graphs, it is difficult to apply existing entity alignment approaches that rely on the structural similarity. Considering certain properties such as name, address, and geographic coordinates is challenging due to the incomplete properties and imprecise geolocations. Using the whole data present in the OSM objects and KG entities can enhance the linking performance and create accurate links. Iteratively linking schema elements and entities can further improve the linking performance.

Although linked entities and schema elements help make the geographic data easily accessible to semantic applications, the unlinked data remains inaccessible. Our third research question deals with converting OSM data into a structured semantic knowledge source.

3. How to create a geographic knowledge graph from OpenStreetMap?

OSM's flat and heterogeneous schema can hinder the usage of OSM in semantic applications. Lifting the OSM schema into a hierarchical ontology can help to

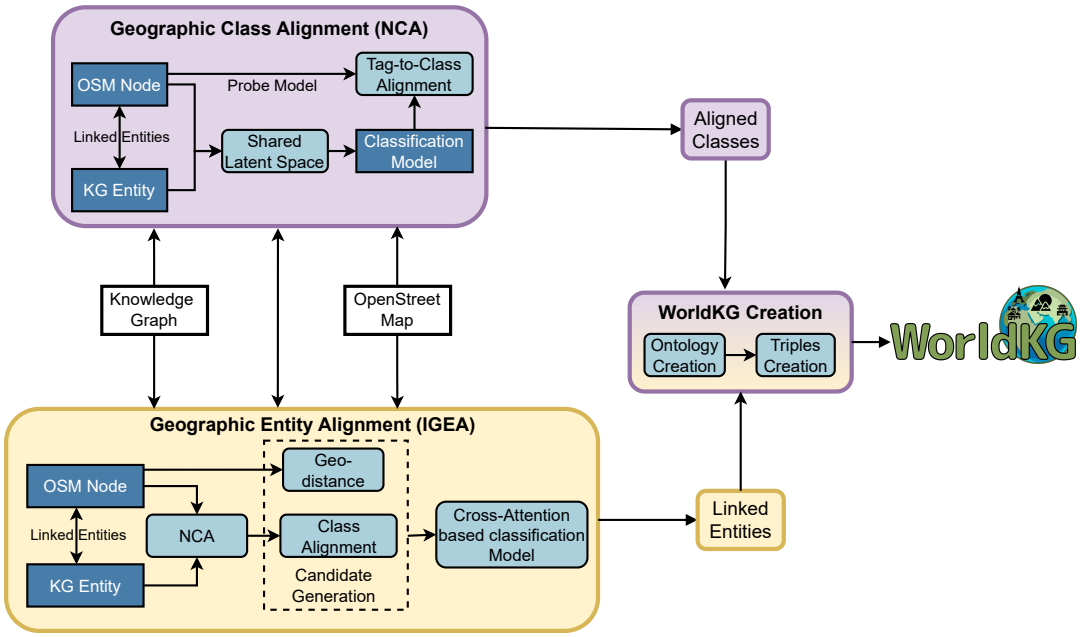


FIGURE 1.2: Overview of the thesis contributions

overcome the challenges of the flatness of the schema. We need to develop a novel ontology that conforms to semantic structure and makes OSM data easily accessible for downstream applications.

1.2 Contributions

To address the research questions designed in Section 1.1, we have proposed several solutions as contributions to this thesis. We contribute to schema alignment, entity alignment, and knowledge graph creation. We make use of various sources of geographic information to design these solutions. Figure 1.2 depicts the overview of the contributions along with their interconnections.

Geographic Schema Alignment: As shown in the upper left part of Figure 1.2, we utilize linked entities from OSM and knowledge graphs to align OSM tags to knowledge graph classes by creating a shared latent space. In the shared latent space, the entities that belong to the same classes are kept in closer proximity. We then probe the model to get the final tag-to-class alignment.

In particular, our contributions include NCA — a novel approach to link class elements between OSM and knowledge graphs. We also propose a shared latent space that combines feature spaces of OSM and knowledge graphs. We propose a novel and effective algorithm to obtain tag-to-class matches from the trained model.

Geographic Entity Alignment: As explained earlier, schema alignment and entity alignment can enhance the performance of each other. Considering this principle, we utilize an iterative approach to align schema and entities. As shown in the lower left part of Figure 1.2, we initially start with the already linked entities from OSM and KG, then apply tag-to-class alignment. We then use tag-to-class alignment along with the geographic distance for candidate generation. Next, we train an attention-based classification model, which classifies pairs of entities into a match or no match. Instead of relying on only certain tags and properties, we consider all properties and tags of entities to fully utilize the rich semantics present in the data.

Overall, our contributions are as follows: Firstly, we propose IGEA, a novel iterative cross-attention-based approach that links the geographic entities between OSM and knowledge graphs. The iterative approach enables the use of tag-to-class alignment and entity alignment simultaneously to improve the candidate blocking and to overcome the annotations and linking sparsity.

Geographic Knowledge Graph Creation: Considering the limitations inherent in OSM's flat and heterogeneous schema, direct access to geographic information within OSM by semantic applications is hindered. To address this issue, we introduce a novel WorldKG ontology designed to transform the flat OSM schema into a hierarchical structure of classes. Additionally, we leverage tag-to-class alignment techniques to establish connections between OSM tags and KG classes, as shown in the right part of Figure 1.2. Subsequently, we construct the WorldKG knowledge graph that conforms to WorldKG ontology.

To summarize, our contributions are as follows: We create the WorldKG knowledge graph that semantically represents data extracted from OSM. Along with the knowledge graph, we present WorldKG ontology that describes the superclass-subclass relations between the class elements from OSM and also provides links to the Wikidata and DBpedia ontology elements.

1.3 Thesis Structure

This thesis is organized into the following sections.

- In Chapter 2, foundational concepts and relevant background for this thesis are described.
- Chapter 3 presents an overview of recent advances in the fields of entity and schema alignments and geographic knowledge graphs.

The subsequent three chapters (4, 5, and 6) provide concise summaries of the published papers, along with my contributions to them, which are attached in the appendix.

- Chapter 4 offers a summary of the paper "Towards Neural Schema Alignment for OpenStreetMap and Knowledge Graphs", addressing the first research question.
- Chapter 5 presents a summary of the paper "Iterative Geographic Entity Alignment with Cross-Attention", addressing the second research question.
- In Chapter 6, a summary is provided for the paper "WorldKG: A World-Scale Geographic Knowledge Graph", addressing the third research question.
- Finally, the thesis is concluded in Chapter 7, outlining the overall results obtained and providing an outlook for future research.

Chapter 2

Background

In this chapter, we present the background necessary to understand further chapters. First, we describe the concept of geographic data along with descriptions of volunteered geographic information datasets. Then, we describe knowledge graphs, their storage, and their querying. Lastly, we present prominent algorithms pertinent to train the volunteered geographic information.

2.1 Volunteered Geographic Information

Geographic data consists of objects/entities that can be located on the earth's surface, ideally with latitude and longitude. In the past years, many smart city applications have emerged which rely on geographic data. The collection and maintenance of such data can be expensive. In 2007, Goodchild [Goo07] invented the term volunteered geographic information (VGI) which includes tools to collect geographic data on the web using volunteered efforts from contributors. The collected data is then stored in a database or file system and can be openly accessed by individuals on the internet. The popularity of VGI is attributed to its open and free nature, since often in many regions, VGI is the only source of geo-information [NZ14]. Currently, there exist many VGI platforms in various data formats that cater to a vast user base. One such data source is OpenStreetMap. Next, we will take a look at OSM in detail.

2.1.1 OpenStreetMap

The OpenStreetMap (OSM) project was initiated in 2004 by Steve Coast at the University of London, which still hosts many OSM data services. Though the project started with the goal of building a global map, the current motivation of the OSM is to provide an openly available geographic resource that can be further explored by routing or navigational applications. Currently, OSM is considered one of the largest openly available sources of volunteered geographic information [NZ12]. The data is made available under the Open Database License (ODbL) [1]. The users can access the OSM data in multiple ways, including a web interface[2] as shown in the Figure 2.1 and an API[3] [4].

Over the past decade, OSM has grown in terms of contributors as well as contributions. It started with 1000 users in 2006 and as of 2024, has over 11 million users. Similarly, in terms of the GPS points recorded, the number increased from 1 million to over 28 billion in past years. Figure 2.2 gives an overview of the growth of OSM

[1] https://opendatacommons.org/licenses/odbl/
[2] https://www.openstreetmap.org/
[3] https://api.openstreetmap.org/
[4] https://www.openstreetmap.org/copyright

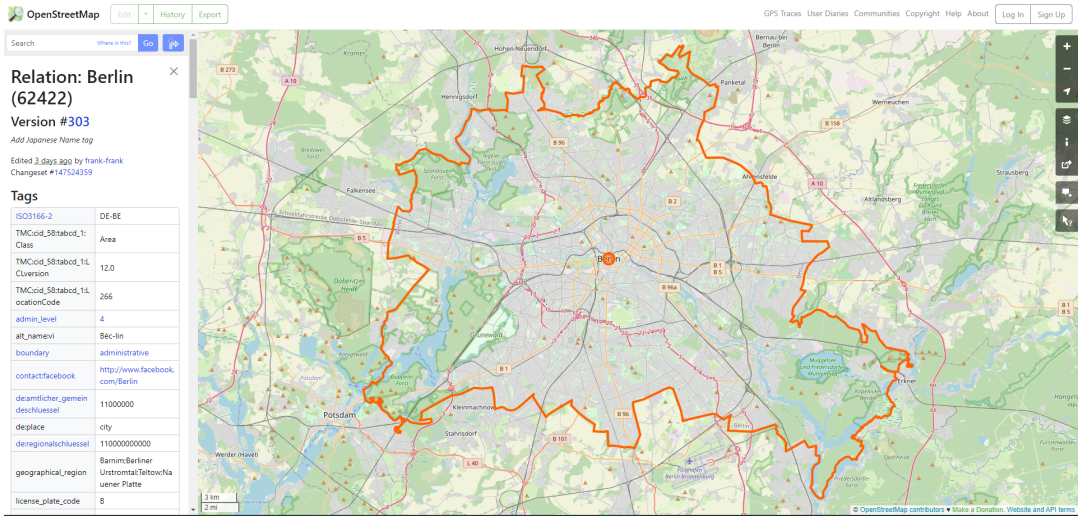
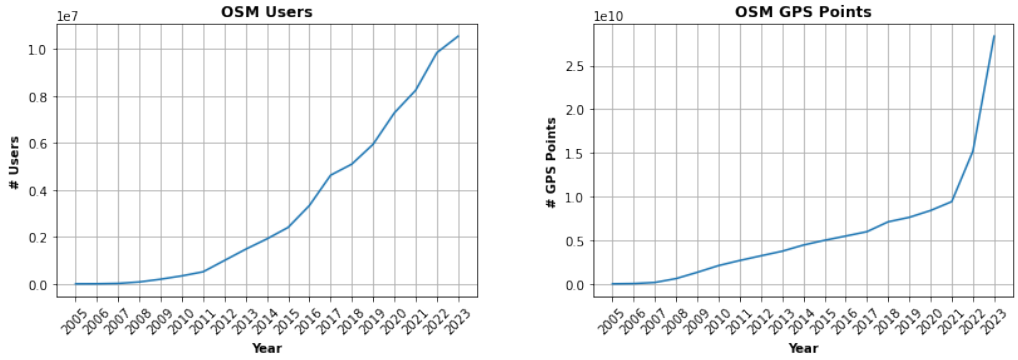


FIGURE 2.1: OpenStreetMap web interface view for the city of Berlin, Germany, ©OpenStreetMap contributors, ODbL



A) Increment in OSM users

B) Increment in OSM GPS points

FIGURE 2.2: An incremental growth of registered OSM users and the number of GPS points collected by contributors

from its inception in 2006 until October 2023 in terms of the users (2.2a) and GPS points tracked (2.2b)[5]. Note that the users reported in the figure are all registered users and not active contributors. The number of active contributors can be less than the number of registered users. The pace of information updates is also high. On 19th February 2024, between 12:00 to 13:00 CST, overall 730 contributors made 288,759 map edits in 117 countries[6]. OSM follows its own schema and data model, consisting of various components.

2.1.2 OSM Data Model

The OSM data model contains various components.

- Each entity in OSM is assigned a unique identifier.
- There are three types of entities described in OSM, namely, nodes, ways, and relations.

[5] https://osmstats.neis-one.org/

[6] https://osmstats.neis-one.org/?item=trending

  - Nodes denote the points containing one pair of latitude and longitude. Examples of node data types are trees, mountain peaks, etc.
  - Ways represent line strings containing an array of latitude and longitude pairs. This data type is used to represent objects such as roads, and rivers. If the way's first node and the last node are the same, then it is called a closed way and is used to describe areas such as buildings or forests.
  - Relations are used to describe complex objects that are created using multiple other data types such as nodes, ways, or relations themselves as subrelations.

- Each OSM object consists of a set of key=value pairs called tags. These tags describe the features of the OSM object. Some key=value pairs describe the type of the given object. For example, in Listing 2.1, *place=city* and *admin_level=4* define the type of the object Berlin.

- Each object also contains a set of metadata that defines the version number, changeset, comments about the current change along with a timestamp.

Formally, we define the OSM object as follows,

Definition 1 (OSM Object) *An OSM object* $\mathcal{O} = (i, nwr, l_{vec}, tags, meta)$ *consists of an id* i*, a type of the object describing if it is a node, way or a relation* nwr*, a point location or an array of members depending on the type of the object* l_{vec}*, an array of key=value pairs tags and a list of the metadata information meta.*

In the Listing 2.1, we show the OSM data for object Berlin in relation form.

OpenStreetMap, although huge, does not provide geographic data in a contextual form that can be utilized for downstream applications. Having OSM data in a structured and contextual format can benefit downstream tasks. In the following, we take a look at the semantic web and its various components and technologies that can be beneficial for the semantic representation of OSM.

2.2 Semantic Web

The semantic web extends the World Wide Web to incorporate a web of data where information is structured and linked in a way that enables machines to understand and interpret its meaning. The semantic web follows the principles of linked data and semantic technologies. Linked data refers to ways of publishing data on the web and linking data within sources for the creation of a vast network of interlinked data from disparate sources [HB11].

The semantic web utilizes standardized protocols such as RDF (Resource description framework) and SPARQL (SPARQL Protocol and RDF Query Language) to create a global web of data. The semantic web addresses the limitations of the traditional web by enriching web contents with machine-readable metadata. As a result, it streamlines processes such as search, discovery, integration, and utilization of data across diverse domains and applications.

2.2.1 Resource Description Framework

Resource description framework (RDF) [Sch+] includes the concepts and notions to describe and represent data on the web. The RDF data model consists of an RDF

```
id                              62422
nwr                             relation

lvec
172 members
member                           id
-----------------------------------------------
Node                             240109189
Way                              50291800
Way                              77913336
Way                              29413660
.
.
.
Way                              506304229

tags
key                              value
-----------------------------------------------------
ISO3166-2                        DE-BE
TMC:cid_58:tabcd_1:Class         Area
boundary                         administrative
admin_level                      4
name                             Berlin
place                            city
.
.
.
website                          http://www.berlin.de
wikidata                         Q64
wikipedia                        de:Berlin

Part of
2 relations
type                            id
-------------------------------------------------
Relation                        8365411
Relation                        8365511

meta
Version                         303
comment                         Add Japanese Name tag
Changeset                       147524359
timestamp                       2024-16-02T11:32:08Z
```

LISTING 2.1: Example of an OSM relation object for the city of Berlin

graph, which is curated using a so-called triple of the form *Subject - Predicate - Object*. In the triple form, the subject and the object are considered nodes and the predicate is

a relation joining the two nodes. The triples are generated by utilizing fundamental components of the RDF model, namely IRIs (Internationalized Resource Identifier), and literals [Sch+].

- IRIs: An IRI (International Resource Identifier) is a Unicode string that follows a set of rules and syntax[7].
- Literals: A literal is used to describe strings, numbers, or dates and can only be placed in the object place.

Another important concept when working with RDF datasets is RDF vocabulary, which consists of a list of IRIs used in the RDF graph. The *namespace IRI* denotes a common substring with which an IRI begins. Namespaces that are commonly used are stored as *namespace prefixes*. Table 2.1 shows some common namespace prefixes and their IRIs.

TABLE 2.1: Common namespace prefixes and their corresponding IRIs

Namespace Prefix	**Namespace IRI**
rdf	http://www.w3.org/1999/02/22-rdf-syntax-ns#
rdfs	http://www.w3.org/2000/01/rdf-schema#
xsd	http://www.w3.org/2001/XMLSchema#
owl	http://www.w3.org/2002/07/owl#

The RDF dataset (collection of multiple graphs) or a graph that follows the strict RDF syntax is called an RDF document. They can have various formats to make it easier to exchange them over the web. Some common formats include Turtle (.ttl), JSON-LD (.jsonld), Notation3 (.n3), and N-Triples (.nt). The listing 2.2 shows an example of entity Berlin from the DBpedia database in Turtle format. Here, various ontologies are used, such as DBpedia ontology, GeoSPARQL ontology, and owl ontology. The example of interlinking of identical elements that refer to the same real-world entity between sources is given by *owl:sameAs* predicate. These links are also called identity links.

2.2.2 Knowledge Graphs

Recently, knowledge graphs have emerged as a source for organizing information in structured formats. Although the concept of using nodes (concepts) and edges (relations) has been in practice since the beginning of the semantic web in the 1990s, the actual term 'knowledge graph' was introduced by Google in 2012 with the creation of Google knowledge graph [Ste+12]. In recent years, due to the advancements in artificial intelligence, natural language processing, and graph database technologies, knowledge graphs have been widely used and adopted by major companies for various application scenarios [Noy+19]. Knowledge graphs follow the RDF data model to structure the data. Each entity is created as a node and its relations are depicted as edges.

Definition 2 [Dso+21] formally defines a knowledge graph.

Definition 2 (Knowledge Graph) *A* knowledge graph $\mathcal{KG} = (E, C, P, L, T)$ *consists of a set of entities* E*, a set of classes* $C \subset E$*, a set of properties* P*, a set of literals* L *and a set of relations* $T \subseteq E \times P \times (E \cup L)$*.*

[7] https://www.ietf.org/rfc/rfc3987.txt

```
@prefix dbo: <http://dbpedia.org/ontology/> .
@prefix dbr: <http://dbpedia.org/resource/> .
@prefix geo: <http://www.opengis.net/ont/geosparql#> .

dbr:Berlin a dbo:City;
   rdfs:Berlin "Berlin" ;
   dbo:areaCode 030 ;
   dbo:country dbr:Germany ;
   dbo:isoCodeRegion "DE-BE" ;
   dbo:populationDensity 4126.00^^xsd:double;
   dbo:populationTotal 3677472^^xsd:nonNegativeInteger ;
   owl:sameAs https://www.wikidata.org/wiki/Q64 ;
   geo:geometry POINT(13.404999732971 52.520000457764) ;
   geo:lat 52.520000^^xsd:float ;
   geo:long 13.405000^^xsd:float .
```

LISTING 2.2: RDF Triples in the Turtle format for an example entity Berlin in DBpedia dataset.

The entities in set E are the real-world entities and classes. P represents the set of properties that connect two entities or an entity to the literal value. Each entity in E belongs to a class C. One entity can belong to one or multiple classes. L is the set of literals that can be used to describe string values, dates, or numerical values. T is the collection of triples of either <*entity - relation - entity* > or <*entity - relation - literal*> form.

Figure 2.3 shows a simple knowledge graph for the city of Berlin. The rounded rectangles show entities, the arrows are relations and the rectangles depict the literals.

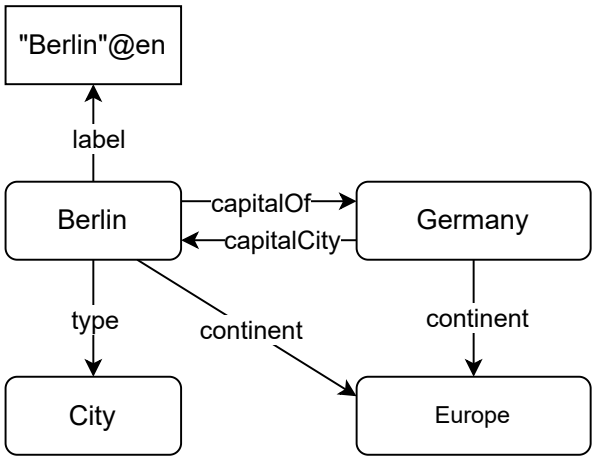


FIGURE 2.3: Example of a simple knowledge graph for the city of Berlin

Currently, many knowledge graphs capture general knowledge and also domain-specific knowledge. These knowledge graphs can be created automatically, semi-automatically, or manually by experts. The most commonly used general-purpose knowledge graphs are Wikidata, DBpedia, and Google knowledge

graph. Figure 2.4 shows the excerpt of Berlin's knowledge representation in the Wikidata knowledge graph. Here, the ids such as *Q64, Q183* are used as entity IRIs (https://www.wikidata.org/wiki/Q64). The *instance of* property is used to describe the type of entity, for example, Berlin is of type *big city* and *federated state of Germany*.

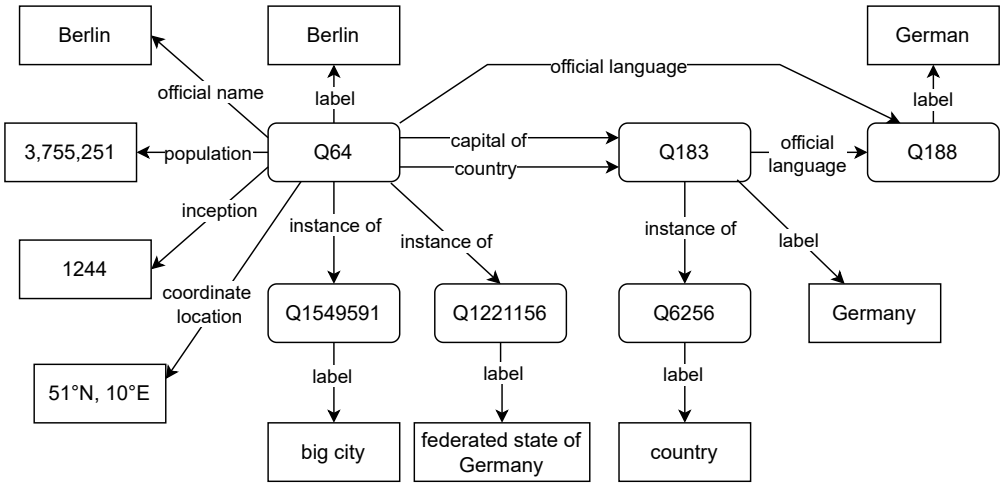


FIGURE 2.4: Wikidata knowledge graph excerpt for the city of Berlin

2.2.3 Geographic Knowledge Graph

In this thesis, we refer to entities of knowledge graph $\mathcal{KG}$ with geo-coordinates L_{geo} are referred to as geographic entities E_{geo}. The knowledge graph created using such entities is called a geographic knowledge graph. Lately, geographic knowledge graphs have been utilized for domain-specific tasks such as geographic question answering [Pun+18; YJY24], Next location recommendation systems [Che+22; Oun+21; PT21; LLS16]. Many of the current knowledge graphs [KMK19; Dso+21] conform to the Open Geospatial Consortium's GeoSPARQL ontology[8]. The geographic objects inside knowledge graphs are defined as *geo:spatilaObject* and have properties such as *geo:hasLength, geo:hasSize*. The *geo:Feature* class has a property named *geo:hasGeometry* which defines the geometry of the spatial object. GeoSPARQL ontology supports point, line, polygon, and multi-polygon representations. Listing 2.3 shows the creation of a geographic Point object using GeoSPARQL ontology along with geographic properties.

2.2.4 SPARQL and GeoSPARQL

SPARQL is a query language designed for RDF data models. SPARQL provides specifications and protocols to query and edit RDF graph content on the Web or in an RDF store. SPARQL provides an extensive syntax for querying RDF data that enables users to retrieve specific information, perform pattern matching, and execute aggregate functions over RDF datasets. It follows a similar syntax to that of an SQL query language. Listing 2.4 queries the graph given in Listing 2.2 to retrieve the country and the population of the city of Berlin.

[8] https://opengeospatial.github.io/ogc-geosparql/geosparql11/geo.ttl

```
@prefix geo: <http://www.opengis.net/ont/geosparql#> .
@prefix qudt: <http://qudt.org/schema/qudt/> .
@prefix rdfs: http://www.w3.org/2000/01/rdf-schema\#
@prefix unit: <http://qudt.org/vocab/unit/> .

eg:berlin
  rdfs:label "Berlin" ;
  rdfs:seeAlso "https://www.wikidata.org/wiki/Q64"
               ^^xsd:anyURI ;
  geo:hasArea [
  qudt:numericValue "891.12"^^xsd:float ;
  qudt:unit unit:KiloM2 ;
  ];
  geo:hasGeometry eg:berlin-geo ;
.

eg:berlin-geo
  a geo:Geometry ;
  geo:asWKT "Point (52.516667,13.383333)"^^geo:wktLiteral ;
.
```

LISTING 2.3: Example of a geographic entity using geoSPARQL ontology

```
SELECT ?city ?country ?population
WHERE {
  ?city rdfs:label "Berlin".
  ?city dbo:country ?country.
  ?city dbo:populationTotal ?population
}

Result:
?city              ?country           ?population
dbr:Berlin         dbr:Germany        3677472
```

LISTING 2.4: SPARQL query to get the country and population for the city of Berlin

SPARQL comprises various functionalities that make it easier to extract the data from an RDF graph. The *Filter* keyword ensures that only patterns for which the filter expression is TRUE are returned. Similarly, the keyword *Optional* is used when the query data may or may not be present in the result set. *Distinct* keyword only displays unique results. The *Order by* keyword is used to order the result set in ascending or descending order.

GeoSPARQL extends SPARQL to add support for geospatial data. As explained in Section 2.2.3, the geographic objects can be represented using multiple geometries and geometric features. GeoSPARQL introduced functions such as *sfContains, sfWithin, sfTouches* to retrieve and manipulate geographic data.

2.3 Machine Learning Algorithms for Alignment of Geographic Sources

The sources of volunteered geographic information are sparse, heterogeneous, and incomplete. As explained earlier, aligning these sources at the schema and entity level can benefit the downstream applications. Since it is a challenging task, we rely on recent developments in the field of machine and deep learning for the alignment. In this section, we describe the machine and deep learning methods that we have utilized to train the models.

2.3.1 Problem Definition

In this thesis, we work on two alignment tasks: geographic entity and schema alignment, especially, tag-to-class alignment. Definitions 3 and 4 [Dso+23]define the tasks of the geographic entity and schema alignment.

In geographic entity alignment, we aim to align entities that represent the same real-world entity. From the example shown in Table 1.1, *Mount Everest* from OSM will be linked to *Mount Everest (Q513)* from Wikidata.

Definition 3 (Geographic Entity Alignment) *Given an entity n from a geographic data source $\mathcal{C}$ ($n \in C$), and a set of geographic entities E_{geo} from a knowledge graph $\mathcal{KG}$, $E_{geo} \subseteq KG$, determine the entity $e \in E_{geo}$ such that $sameAs(n, e)$ holds.*

Here, the sameAs function refers to the predicate of owl ontology[9] which maps two same real-world entities.

In geographic class alignment, we align the schema elements that represent the same real-world concept. In the example in Table 1.1, the OSM tag *natural=peak* will be aligned to the class *mountain* from Wikidata.

Definition 4 (Geographic Class Alignment) *Given a geographic data source $\mathcal{C}$ and a knowledge graph $\mathcal{KG}$, find a set of pairs of class elements of both sources, such that elements in each pair (s_i, s_j), $s_i \in C$ and $s_j \in KG$, describe the same real-world concept.*

2.3.2 Alignment Methods and Models

Alignment methods mainly rely on structural or textual features. Since OSM does not have a fixed ontology, in this thesis we mainly rely on textual features to align elements. In this section, we describe the fundamentals of the training of a model with textual features.

Feature Representation

In our approaches, we utilize OSM tags (key=values) and knowledge graph properties as our features. We form a corpus using the features and treat them as textual inputs. Using the whole set of tags and properties from OSM and knowledge graphs as features can create an extremely sparse feature set. There exist feature representation methods [MRS08] such as bag-of-words, and TF-IDF where the texts are represented either in a boolean way or with simple computations depending on the frequency of occurrence. In the past decade, word embedding methods such

[9] https://www.w3.org/TR/owl-ref/#sameAs-def

as word2vec [Mik+13], glove [PSM14], fastText [Boj+17] have been extensively used for textual feature representation. In 2018, language models such as BERT [Dev+19] became popular for text encoding and representation. BERT is a transformer model. It is trained on a deep bidirectional representation of text and helps in conditioning left and right context, where the context of a word is maintained in overall input. In the above models, the embeddings of a single word are given. In 2019, Reimers et al. [RG19], presented a modification to BERT called sentence-Bert using Siamese and triplet models to get the contextual sentence embeddings. In our approaches, we have utilized, TF-IDF, fastText, and sentence-Bert for feature representation.

Traditional Machine Learning Models

Traditional machine learning models for classification have been utilized for alignment tasks in the past [MBR01; NA11]. These algorithms include decision trees and random forest trees. These models rely on precalculated features such as string similarity and other lexical features of the entity pair. In the case of the alignment tasks where both sources follow different schema and structure, applying simple machine learning algorithms do not yield good results as they only utilize certain features such as names and do not consider the structural and schematic differences.

Neural Networks

Neural networks [Gur97] are computational models inspired by the human brain that are composed of interconnected nodes. They use weighted connections and activation functions to learn complex patterns in the data. The selection of hyperparameters, such as learning rate, number of layers, and activation functions, plays a crucial role in determining the network's performance and generalization ability across various tasks. In the context of textual data, word embeddings are often used as input features and are passed through multiple hidden layers of the neural network. There exist various networks such as simple feedforward, convolutional, recurrent, and long short-term memory models.

In the feedforward network [BG94], the information flows in one direction from the input nodes to the output nodes through hidden layers. The output of the neuron is the weighted sum of its inputs plus the bias. We then apply an activation function to get the final output. Convolutional networks [Gu+18] are widely utilized in the image domain as they can recognize the visual patterns from the images. Each neuron in the convolutional layer is connected to a small subset of input neurons. A filter known as kernel is applied to the input feature to generate an output feature map, on which a non-linear activation function is applied.

When working with sequential data, simple feedforward networks cannot capture the long-term dependencies, as they do not have memory units. Recurrent networks [PMB13] have connections within the layers that allow information to persist over time. Each neuron has input from the current state as well as output from the past state to maintain the sequence. Although recurrent networks are powerful for sequential data, they suffer from the problem of vanishing gradient where, as the network progresses, the gradient gets negligible, making it difficult to learn the long dependencies. Long short-term memory networks [Sai+15] overcome the vanishing gradient problem by utilizing various gated units such as input gate, forget gate, and output gate. The input gate decides how much new information should be stored in a cell state. The forget gate determines how much information from the previous cell

is to be retained in the current cell state. The output gate controls how much of the cell state to be exposed to the next hidden state. These gates work together to regulate the flow of information through the LSTM cell, enabling it to capture long-term dependencies in sequential data.

In neural networks, the intermediate layers form latent spaces that can be utilized for better model understanding and intermediate results generation.

Feature Space Alignment

When working with multiple data sources, neural networks with different settings can be jointly learned to extract shared features and create shared latent spaces. Development of such models is dominant in the fields of computer vision [Fer+13] and machine translation [Lam+18]. One of the approaches proposed by Ganin et al. [Gan+16] aligns the source and target domain by using a neural domain adaptation algorithm that trains a model by taking labeled data from a source domain and unlabeled data from a target domain. While this approach aligns similar distributions of feature spaces, the gradient reversal layer proposed in [Gan+16] can be utilized to form joint spaces based on similarities or differences. In our work, we utilize the gradient reversal layer to form the shared latent space between OSM and knowledge graphs. Another network that utilizes joint learning is a recently published transformer network based on attention mechanisms.

Attention Mechanism

The attention mechanism [Vas+17] trains the deep learning model by selecting important features that help improve the efficiency and accuracy of the model. The attention function consists of key K, value V, and query Q and then mapped to the output.

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V, \tag{2.1}$$

where d_k is the dimension of the key vector [Vas+17].

Attentions are of two types, additive and multiplicative. When attention is applied to the same input sequence, it is known as self-attention. Attention applied to two different inputs is known as cross-attention. Cross-attention has been proven effective when working with multi-modal inputs [Wei+20]. It is modeled similar to self-attention but instead of using only one input as K, Q, and V, it uses a combination of inputs from two different sources as K, Q, and V, the selection of which varies according to the application scenario. In our work, we consider OSM and knowledge graphs as two different modes of data and apply cross-attention to the OSM and KG features.

2.3.3 Evaluation Metrics for Alignment

Alignment problems can be evaluated in multiple ways. One of the ways is for each entity, the output is a list of ranked probable matches, and they are evaluated based on which position the correct match is. In this work, we consider the alignment problem as a classification problem, wherein we classify a pair of entities into

a match or no-match class. We evaluate the alignment problem using classical classification evaluation measures such as precision, recall, and f1-measure. Following, we describe these metrics.

Let us assume that, T is the total number of true pairs present in the ground truth, t_r is the total number of pairs identified by the model and t_p is the number of true pairs identified by the model.

Precision: It is defined as the ratio of all pairs that are correctly identified by the model to all the pairs in the result set.

$$Precision = \frac{t_p}{t_r} \tag{2.2}$$

Recall: It is defined as the ratio of all pairs that are correctly identified by the model to all the pairs in the ground truth.

$$Recall = \frac{t_p}{T} \tag{2.3}$$

F1-Score: It is calculated as a harmonic mean between precision and recall.

$$F1 - Score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{2.4}$$

Having lower values of precision or recall can give an overall lower score to the model. F1-Score balances the precision and recall and has proven to be beneficial in the case of class imbalance [Der16]. In terms of alignment tasks, the data is generally imbalanced as we have many candidate pairs out of which only one pair is the true pair. Using only precision or recall may not reflect the model's capabilities to the full extent. We use F1-Score as the measure to evaluate the performance of our models.

Chapter 3

Literature Review

In this chapter, we offer a comprehensive review of advancements in both schema and entity alignment. Furthermore, we delve into recent developments specifically focusing on geographic entity and schema alignment, highlighting notable contributions. Lastly, we provide an overview of the current advances in terms of the geographic knowledge graphs.

3.1 Schema Alignment

Schema alignment refers to the task of aligning schema elements such as classes and properties. In this section, we discuss the state-of-the-art approaches for schema alignment, ontology alignment, and tabular data alignment. We stick to general-purpose schema alignment approaches, as there is not a research branch dedicated to schema alignment approaches for geographic data sources.

3.1.1 Ontology Alignment

Ontology alignment, sometimes referred to as ontology matching, links the same real-world concepts from various ontologies. There exist several benchmark ontology alignment approaches due to the W3C SWEO Linking Open Data community project[1] and the Ontology Alignment Evaluation Initiative (OAEI)[2] [Alg+19]. The alignment can be carried out at the structure or element level. The element-level ontology alignment approaches use intrinsic features such as names or other descriptive properties and apply string similarity measures on the pairs of elements. Generally, fussy string matching approaches such as Jaro-Winkler, and Levenshtein distances are used to account for the spelling variations or mistakes [Li+09]. On the other hand, the structural level approaches rely on the structural similarities between elements. This includes using the ancestors and descendants along with the neighbors to calculate the similarity [NBT13]. Relying on string similarity or structural similarity has proven less effective and newer approaches have tried to incorporate semantic information from various sources such as WordNet, and Wikipedia to align ontologies [Jai+10]. Although, effective use of both string and semantic similarity is still a challenge [ORG15].

To overcome the challenges of similarity-based approaches, many approaches have adopted machine learning techniques for ontology alignment. The GLUE architecture [Doa+04] presented multiple techniques to learn the semantic mappings

[1] https://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData
[2] OAEI evaluation campaigns: http://oaei.ontologymatching.org

in a semi-automatic way. Nkisi-Orji et al. [Nki+18] proposed an approach that incorporated semantic similarity using word embeddings and string similarity. The model was then trained using a random forest algorithm to get the final matches. More recently, deep learning-based approaches have gained popularity for ontology matching. Bento et al. [BZG20] proposed an architecture with convolutional neural networks where string matching is performed using character embeddings. They also included class hierarchy for optimal results. An unsupervised representation learning-based approach [Qiu+17] used correlations between different entity descriptions to learn representations of entities, which were used to get matching ontology elements using supervision. In ERSOM [Xia+15], stacked auto-encoders were used for higher-level description learning. Later, the iterative similarity propagation method was used to get the alignment.

3.1.2 Tabular Data Schema Alignment

Tabular data schema alignment refers to aligning schema elements of data, such as relational databases [RB01]. The EmbDi [CPT20] approach created graphs based on tabular data. This graph structure consisted of schema elements and entities. EmbDi generated sentences from the graph to get the embeddings that are used to find similarities between the schema elements. Rema [Kou+20] trained embeddings using random walks generated from graph relations to get the column mappings. Madhavan et al. presented the Cupid [MBR01] approach that matches schema elements based on element names, structure, constraints, and data types. Cupid uses linguistic, structural, and contextual matching to find the mapping between schema elements. Similarity Flooding [MGR02] transformed a table into a directed labeled graph in which nodes represent columns to compute similarity values iteratively.

OSM and knowledge graphs present structural and schematic differences, hence direct application of aforementioned ontology alignment and tabular data schema alignment methods is not feasible on the task of OSM to KG schema alignment. In our approach, we utilize entity descriptions along with state-of-the-art neural methods to align schema elements

3.2 Entity Alignment

Entity alignment aims to align entities across different sources that refer to the same real-world object. In linked data, aligned entities are represented using *owl:sameAs* link. In this section, we focus on the related work in the field of entity alignment and geographic entity alignment.

3.2.1 Generic Entity Alignment

There has been advancement in the field of entity alignment for knowledge bases and graphs. Similar to schema matching, these methods include string and structural similarity-based approaches as well as machine learning and deep learning-based approaches. LIMES framework [SNL17] utilized several rules to get the candidate pairs and then used these rules to train the classifier which predicts the links. LIMES contains multiple algorithms such as Eagle [NL12], Coala [NLC13], Euclid [NL13], and Wombat [SNL17] based on unsupervised and active learning. Silk [Vol+09] framework introduced a link specification language that specifies heuristics

to test whether a *sameAs* link exists between entities. Hao et al. [Hao+16] proposed a joint learning model by only using the structural information to align multilingual knowledge bases. Recently, deep learning-based models have gained popularity for the task of entity alignment on tabular data. DeepMatcher [Fu+20] and HierMatcher [Mud+18] used an embedding-based deep learning approach for predicting the matches for tabular datasets. Peeters et al. [PB22] used contrastive learning with supervision to match entities in small tabular product datasets.

3.2.2 Geographic Entity Alignment

Geographic entity alignment aims to align geographic entities across different geographic sources that refer to the same real-world object. In the past, approaches often relied on geographic distance and linguistic similarity between the labels of the entities. LGD approach [ALH09] used a quadratic function for spatial distance along with the string similarity to obtain links between entities. Karalis et al. [KMK19] utilized Jaro–Winkler string similarity and geographic distance between entities to get entity alignments. Tempelmeier et al. [TD21] proposed the OSM2KG algorithm – a machine-learning model to learn a latent representation of OSM nodes and align them with knowledge graphs. OSM2KG also used KG features such as name, popularity, and entity type to produce more precise links. LIMES framework introduced ORCHID [Ngo13] and Radon [She+17] as link discovery algorithms for geospatial data. These algorithms work with the polygons and DE-9IM relations to align entities.

In this thesis, we build on top of existing methods and use state-of-the-art deep learning algorithms to get accurate schema and entity matches.

3.3 Geographic Knowledge Graphs

Knowledge graphs have gained popularity over the past decade and have been used in many semantic applications. In this section, we reflect on the current advances in the field of geographic knowledge graphs. As mentioned in Section 2.2.3, we consider knowledge graphs or subset of knowledge graphs where for an entity a geographic location is present as geographic knowledge graphs. Table 3.1, provides an overview of the knowledge graphs with geographic data.

TABLE 3.1: Overview of the current geographic knowledge graphs

Name	Ontology	Sources	Scope
Wikidata [VK14]	Wikidata	Multiple	World
DBpedia [Leh+15]	geoSPARQL	Multiple	World
LinkedGeoData [ALH09]	LGD	OSM	World
Yago2Geo [KMK19]	YAGO2	Multiple	GR, UK, IR, US
KnowWhereGraph [Jan+22]	SOSA[3], QUDT[4]	Multiple	US
WorldKG [Dso+21]	WorldKG	OSM	World

Wikidata [VK14], one of the biggest knowledge graphs, contains many entities with geographic locations from all over the world. As of February 2024, there were

[3] https://www.w3.org/TR/vocab-ssn/
[4] http://www.qudt.org/

10,805,736 entities with geographic location property (P625). These locations oftentimes are imported from other sources, which can lead to imprecise coordinates. Wikidata follows its own ontology and has coordinates in radians format as well as with WKT[5] format. The representation is uneven, with some classes such as railway stations being represented more than other classes such as bike charging stations. DBpedia [Leh+15], a knowledge graph curated from Wikipedia, is also a general-purpose knowledge graph that has over 1.2 million geographic entities. DBpedia follows OGCs geoSPARQL[6] ontology to represent geographic entities with *geo:geometry, geo:lat and geo:long* predicates.

LinkedGeoData [ALH09] was one of the first attempts at bringing the OSM data into RDF formats. The ontology is created manually and contains classes such as cities, amenities, and public transport (500 classes). Moreover, the knowledge graph does not ensure that only quality nodes (nodes with at least one tag) from OSM will be lifted into the knowledge graph. On similar lines, Yago2Geo [KMK19] extended YAGO2 [Hof+13] knowledge graph with geographic knowledge. It contains the geographic data that is collected from the Greek Administrative Geography dataset, Ordnance Survey data from Ireland and Northern Ireland, the Global Administrative Areas dataset [GDA12], and OSM. The geometries are introduced using OGC vocabulary. Yago2Geo also contains temporal information. The Yago2 ontology is extended manually to include the geographic data from the aforementioned datasets.

Recently, Janowicz et al. [Jan+22] published an event-centric geographic knowledge graph that accumulates data from various sources and can answer questions such as what happened in this place in the past. Instead of using points or polygons for regions, KnowWhereGraph uses the S2 grid system [BRR20], which covers the earth's surface with hierarchical grids. The graph contains data from over 16 data sources and has 27 different data layers. The graph focuses on domain application scenarios such as climate hazards, wildfires, and air quality.

To overcome the shortcomings of the previously built geographic knowledge graphs, in this thesis, we present WorldKG, a knowledge graph based on OSM data built using a novel WorldKG ontology.

[5] https://docs.ogc.org/is/18-010r7/18-010r7.html
[6] https://www.ogc.org/standards/geosparql

Chapter 4

Towards Neural Schema Alignment for OpenStreetMap and Knowledge Graphs

Publication Details:
Alishiba Dsouza, Nicolas Tempelmeier, and Elena Demidova.
"Towards Neural Schema Alignment for OpenStreetMap and Knowledge Graphs".
In: Proceeding of the 20th International Semantic Web Conference, ISWC 2021. Springer, 2021 pages 56–73, 2021.
DOI: 10.1007/978-3-030-88361-4_4

4.1 Summary

Geographic data sources such as OpenStreetMap (OSM) rely on semi-structured key=value pairs to describe objects. However, these pairs, also known as tags, lack the semantic richness required for direct accessibility by semantic applications. Knowledge graphs (KG), on the other hand, provide precise semantics for their entities but do not have large coverage of geographic information. Integrating OSM and knowledge graphs at the schema level can help in making a wide range of geographic entities from OSM available for semantic applications. Such an alignment can also help OSM volunteers to correctly map an entity, as OSM tags are not always intuitive. For example, an OSM tag *natural=peak* describes a real-world concept called *mountain*. Aligning *natural=peak* tag to Wikidata class *Mountain (Q8502)* can help the OSM volunteers better understand the tags, in turn creating quality entities in OSM. However, achieving such alignment poses challenges due to the schema differences between OSM and knowledge graphs, the flat and sparse nature of the OSM schema, and the absence of pre-existing links between OSM and knowledge graphs.

Past schema alignment approaches that rely on string similarity do not yield good results due to representational differences between entities (e.g., natural=peak and mountain). Approaches that consider structural similarity cannot be applied due to the flat OSM schema. Tabular data alignment approaches do not produce results, since the conversion of OSM data into tabular form leads to spare data. In this thesis, we present **NCA** approach that aligns OSM schema elements called tags to knowledge graph classes. To align schema elements, one can utilize various tasks

such as entity alignment, and node classification. In this work, we utilize the already linked OSM and KG entities, simultaneously classifying these entities into KG classes.

The contributions of this work are as follows:

- We present a novel approach to class alignment for OSM and knowledge graphs.
- We propose a novel shared latent space that fuses feature spaces from knowledge graphs and OSM in a joint model, enabling simultaneous training of the schema alignment model on heterogeneous semantic and geographic sources.
- We develop a novel, effective algorithm to extract tag-to-class alignments from the resulting model.
- The results of our evaluation demonstrate that the NCA approach is highly effective and outperforms the baselines by up to 37 percentage points in terms of F1-Score.

Our NCA approach consists of two steps. In the first step, we use linked OSM and KG entities to build an auxiliary classification model. This model takes as an input the OSM and KG entities along with their features. These entities are then classified into KG classes. To jointly represent OSM and KG entities in a shared space, we use an adversarial classifier. The main goal of using an adversarial classifier is that the model should not be able to distinguish between OSM and KG samples, in turn making a shared latent space that similarly represents both OSM and KG entities. By classifying OSM entities into KG classes, we align the shared latent space in such a way that we have OSM and KG entities that belong to the same classes in closer proximity to each other. Once we have the classification model, we then probe this model with a tag to get the KG class. For each given tag, we carry out a full forward propagation and the class is determined using the activation values of the last layers. We only select the pairs of tag and class if they are above a certain threshold, which is determined experimentally.

We evaluate our model on the various country datasets of OSM, Wikidata, and DBpedia and compare it against state-of-the-art baselines for schema alignment for tabular data and string similarity-based baselines. We evaluate the tag-to-class alignment in terms of precision, recall, and f1-measure. Our experiments indicate that the NCA approach outperforms all baselines on all datasets in terms of the f1-measure. Specifically, we obtain 13% points and 37% points improvement on Wikidata and DBpedia datasets, respectively. Table 4.1 shows some example tag-to-class alignments that are generated using the NCA approach. Our NCA approach can find matches that are not lexically similar, such as *amenity=cinema* and *movie theater*.

TABLE 4.1: Tag-to-class alignments obtained using NCA approach

Wikidata			
France	Germany	Great Britain	USA
amenity=bicycle_rental: bicycle-sharing station	amenity=cinema: movie theater	railway=station: railway station	landuse=reservoir: reservoir
DBpedia			
France	Germany	Great Britain	USA
railway=station: Place	place=municipality: Place	place=hamlet: Place	man_made=lighthouse: Location

We notice that the performance is dependent on the data quality of the number of identity links, class, number of entities per class, and distinct tags and classes. To prove the importance of the shared latent space, we conducted an ablation study wherein we removed the shared-latent space and compared the performance to the NCA model with the shared space, we observed that the performance increased by 34 and 11% points on Wikidata and DBpedia in the F1-Score due to the latent space.

4.2 Contributions

I contributed to conceptualizing and developing the methodology. I primarily handled the major implementations and conducted experiments and evaluations on the approach. Finally, I contributed to the writing and reviewing of the manuscript.

Chapter 5

Iterative Geographic Entity Alignment with Cross-Attention

Publication Details:
Alishiba Dsouza, Ran Yu, Moritz Windoffer, Elena Demidova.
"Iterative Geographic Entity Alignment with Cross-Attention"
In: Proceedings of the 22nd International Semantic Web Conference, ISWC 2023.
Springer, 2021, pages 216–233, 2023.
DOI: 10.1007/978-3-031-47240-4_12
We received the best student paper award for this publication.

5.1 Summary

As discussed in Chapters 1 and 4, current general-purpose knowledge graphs do not have wide coverage of geographic entities. Volunteered geographic information sources such as OSM do contain a huge amount of geographic data, but the heterogeneous and flat schema makes it difficult to utilize the data to its full potential. Although knowledge graphs inherited from OpenStreetMap such as LinkedGeodata [ALH09], Yago2Geo [KMK19] provide machine-readable semantics, they cover a fraction of the classes and entities. These sources do not contain entity-to-entity links that can provide better contextual information. Aligning knowledge graphs and OSM at the entity and schema level can profit both sources and make them more beneficial for downstream applications such as geographic question answering and information retrieval. The interlinking becomes challenging due to numerous reasons. One of the biggest challenges is the sparsity of entity annotations and links between sources. Currently, only about 0.53% of OSM nodes are linked to the Wikidata knowledge graph. Another challenge is the heterogeneity in the entity representations in terms of the schema as well as the actual values of the properties of entities. For example, as shown in Table 1.1, the values of elevations and geo-coordinates differ. In OSM, it is also not straightforward which tag represents the class of an entity.

Over the past years, several approaches such as LinkedGeoData [ALH09], yago2geo [KMK19] have tried aligning these sources at entity and schema level. Many of these approaches have focused on specific administrative regions using well-annotated entities and classes. Some of these approaches rely on a few properties such as the name of the place or the distance and do not consider the rich heterogeneous information present in the entities. They also do not consider the representation heterogeneity and annotation sparsity, which makes the alignment

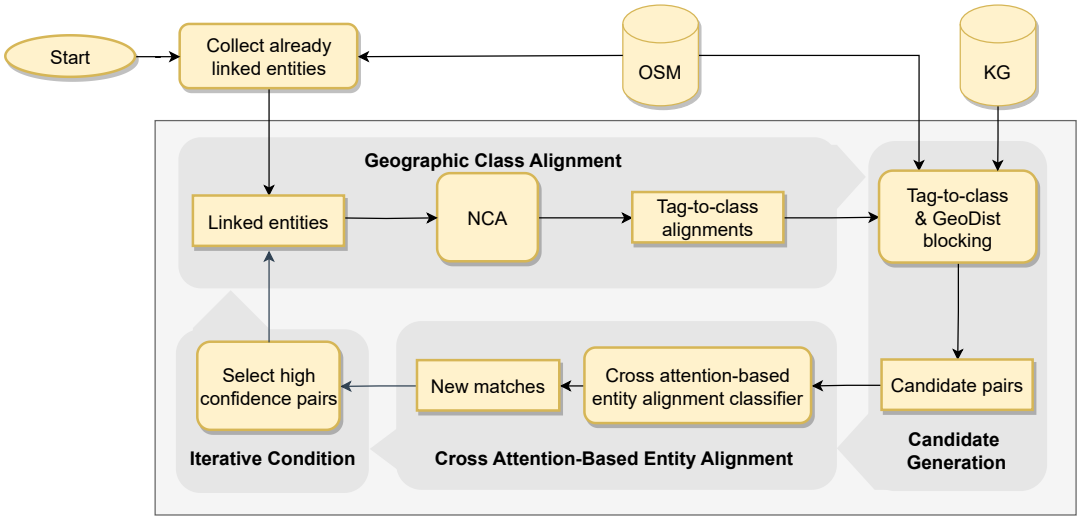


FIGURE 5.1: Overall process for IGEA approach [Dso+23], Copyright ©2023 Dsouza et al.

difficult. Overall, approaches that rely on string similarity or geographic distance while only using certain properties such as names are not sufficient for such alignment.

To overcome the challenges and shortcomings of the current state-of-the-art methods, we propose IGEA – a new iterative approach based on cross-attention to align entities with the heterogeneous contextual representation. The IGEA approach first relies on linked entities and applies class alignment using the NCA [DTD21] approach. The aligned classes along with the geographic distance between entities are then used for candidate blocking. Further, a cross-attention-based module is used to classify a pair of entities into a match or no match. This process is done iteratively. Figure 5.1 shows the overall process for the IGEA approach. Our contributions are as follows:

- We propose IGEA – a novel iterative cross-attention-based approach to interlink geographic entities, bridging the representation differences in community-created geographic data and knowledge graphs.
- To overcome the sparsity of annotations and links, IGEA employs an iterative method for tag-to-class and entity alignment, with integrated candidate blocking mechanisms for efficiency and noise reduction.
- We demonstrate that IGEA substantially outperforms the baselines in F1-Score through experiments on several real-world datasets.

The approach first takes the already linked entities from OSM and knowledge graphs and applies the NCA approach to get the tag-to-class alignment. Once we have the tag-to-class alignments, we then apply candidate generation. Creating a candidate set allows us to reduce the search space and compare entity pairs that are a probable match, as not all OSM entities will have a corresponding KG entity. We employ a two-way candidate blocking. First blocking is entity type-based, wherein we use the tag-to-class matches obtained in the previous step for the class matching since the same real-world entities should belong to the same class. The second filtering is done based on geographic distance. Therefore, if the geographic

distance between two entities is less than 2500 meters [TD21], and they belong to the same class, then they are considered as a candidate pair. Once we have these pairs, we then apply the cross-attention-based entity alignment classification. The introduced cross-attention module helps in understanding the important keys or properties across sources as well as within a given source. We utilize the linked entities as the supervision for the classification, along with the geographic distance between the entities. The classification layer predicts whether the given pair is a match or not. Finally, we create an end-to-end iterative pipeline to align OSM and KG entities, which only selects high-quality pairs using a confidence threshold.

To evaluate the effectiveness of the IGEA approach, we use various datasets curated from various countries from OSM, Wikidata, and DBpedia. We compare the obtained results with state-of-the-art baselines for entity alignment, such as LinkedGeodata [ALH09], Yago2Geo [KMK19], DeepMatcher [Mud+18], OSM2KG [TD21]. IGEA approach outperformed baselines in terms of F1-Score and archives, up to 18% and 14% points improvement over Wikidata and DBpedia datasets. Our experiments regarding the number of iterations showed that, for many datasets, up to the 3rd iteration the performance gradually increases and then stays stagnant after the 4th iteration. This trend was followed for entity alignment as well as tag-to-class alignment. We also conducted an ablation study, which proved the effectiveness of the individual components that were introduced as part of the approach.

5.2 Contributions

I devised the current approach and methodology, incorporating an attention mechanism and additional features. Moritz Windoffer and I implemented the approach, while I conducted experiments and evaluations. I also contributed to the writing and reviewing of the manuscript.

Chapter 6

WorldKG: A World-Scale Geographic Knowledge Graph

Publication Details:
Alishiba Dsouza, Nicolas Tempelmeier, Ran Yu, Simon Gottschalk, Elena Demidova.
"WorldKG: A World-Scale Geographic Knowledge Graph".
In: Proceeding of The 30th ACM International Conference on Information and Knowledge Management, CIKM 2021. ACM, 2021, pages 4475–4484, 2021.
DOI: 10.1145/3459637.3482023

6.1 Summary

As mentioned in Section 1, OSM is semi-structured. Due to its volunteered nature, the created entities do not follow a fixed schema, as volunteers are provided only with a set of guidelines to create the entities. This, along with the heterogeneous and incomplete entity descriptions, creates a bottleneck for downstream machine-learning applications that rely on structured data. Having OSM data in knowledge graph form can overcome the challenges pertaining to structured data and heterogeneity. In the past, approaches such as LinkedGeoData [ALH09], Yago2Geo [KMK19] have attempted to adopt the OSM data into a structured KG ontology. These approaches, however, lack in terms of geographic coverage and represent entities belonging to a few classes. Hence, we need a source of geographical semantic information that overcomes the shortcomings of previously mentioned approaches along with the challenges occurring due to the OSM schema. To this end, we propose WorldKG knowledge graph that provides comprehensive semantic information of geographic entities and is based on OSM nodes. It has over 100 million geographic entities from 188 countries.

Our contributions are as follows:

- We present WorldKG – a new knowledge graph containing large-scale semantic geographic data extracted from OSM.
- We present the WorldKG ontology, which semantically describes geographic entities and links them to the specific classes in the Wikidata and DBpedia ontologies.

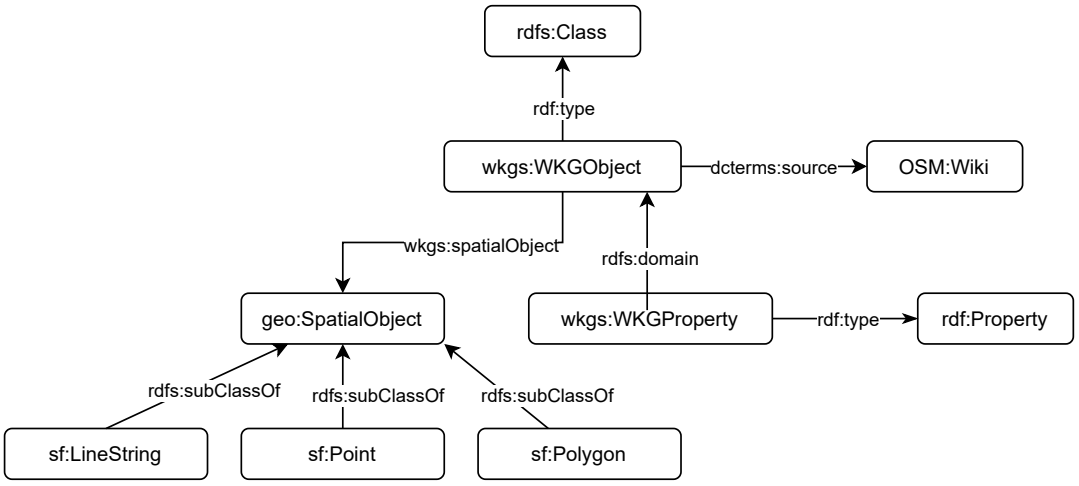


FIGURE 6.1: WorldKG ontology [Dso+21], Copyright ©2021 ACM.

- We provide access to WorldKG through a SPARQL endpoint and provide downloadable data files in the standard RDF turtle format[1].

We create a novel WorldKG ontology by exploiting the keys and values from OSM to build a hierarchical schema. Figure 6.1 shows the snippet of the WorldKG ontology. Each WorldKG entity here named as *WKGObject* has a property of a type *spatialObject* that can either be a point, line-string, or polygon and represents the geometry of the given entity. The *rdfs:Class* defines the type of the *WKGObject*. The *WKGObject* can have multiple properties such as *name*, *adressCountry*. The provenance of each object is ensured through the *OSM:Wiki* property. We build the class hierarchy using the tags of OSM. For example, in tag *amenity=restaurant*, *amenity* becomes the superclass and *restaurant* becomes the subclass. The tag-to-class pairs inferred by our NCA [DTD21] approach are incorporated into the ontology using *Owl:equivalentClass* property.

The creation process is divided into 2 steps. The first step creates the ontology and then in the second step, based on the ontology, we create the triples that form the knowledge graph. As mentioned earlier, OSM does not have a fixed schema, hence, including all the tags and keys from OSM is not feasible. OSM map features[2] provide information regarding tags that possess class information and their attributes. We use these tags as classes in WorldKG ontology. We then produce the hierarchical schema from OSM tags, the keys are considered super-class, and values are considered subclasses.

We convert the names of classes and properties to fit to the OWL naming conventions[3]. Later, we align the WorldKG classes to Wikidata and DBpedia classes that were inferred using the NCA approach. Once we build the ontology, we create triples for any of the OSM snapshots. We only consider OSM nodes that have at least one tag. To ensure quality, we filter out tags and keys based on the WorldKG ontology. The final step is to create the triples and validate them to get the whole WorldKG knowledge graph.

The current version of WorldKG has around 800 million triples of over 113 million entities. 33 superclasses such as *place*, and *amenity* are essentially the keys of

[1] https://www.w3.org/TR/turtle/
[2] https://wiki.openstreetmap.org/wiki/Map_features
[3] https://www.w3.org/TR/owl-ref/

OSM, which are then divided into around 1100 subclasses. We have over 1800 unique properties present in WorldKG. 40 Wikidata and 21 DBpedia classes are linked to WorldKG classes using the NCA approach. Table 6.1 provides an overview of the knowledge graph statistics.

TABLE 6.1: WorldKG knowledge graph statistics [Dso+21], Copyright ©2021 ACM.

Quantity	Count
Total triples	828,550,751
Total entities	113,444,975
Top-level classes	33
Subclasses	1,143
Unique properties	1,820
Links to Wikidata classes	40
Links to DBpedia classes	21

WorldKG has over two orders of magnitude higher geographical entities than Wikidata and DBpedia. To assess the quality of the tag-to-class alignment, we manually investigate a random sample of entities for 10 classes. We observe over 99% accuracy for the tag-to-class alignment. Along with the triples of the knowledge graph, we also provide a SPARQL endpoint based on Virtuoso[4] that enables us to efficiently query the knowledge graph. By making use of the GeoSPARQL [BK11], the SPARQL endpoint can answer geographic queries.

6.2 Contributions

I contributed to the ontology creation and worked on the methodology. Implementations were done by me. I contributed to the quality assurance of the WorldKG. All the authors helped in writing and reviewing the paper.

[4] https://vos.openlinksw.com/owiki/wiki/VOS

Chapter 7

Conclusion

Volunteered geographic information (VGI) sources like OpenStreetMap offer significant potential for leveraging these datasets in downstream machine-learning applications. In its original state, the OSM data is not directly accessible by semantic applications and does not provide needed contextual information. On the other hand, knowledge graphs are the sources of rich contextual information but lack the coverage of geographic information. Integrating VGI sources with sources of semantic information can benefit both sources with rich semantic geographic information. Certain challenges of the OSM's flat and heterogeneous schema and lack of identity links hinder the integration process at the schema and entity levels. We tackle these challenges by developing machine learning algorithms to integrate the OSM data with the knowledge graph and finally present a geographic knowledge graph that adopts OSM data into a semantic source.

7.1 Summary of Contributions

In this section, we summarize our contributions as answers to the research questions presented in Chapter 1. In particular, we contributed to the tasks of geographic schema alignment, geographic entity alignment, and geographic knowledge graph creation.

7.1.1 Geographic Schema Alignment

In Chapter 4, we proposed an approach to solve the problem of geographic schema alignment by aligning tags of OSM with classes of knowledge graphs. Due to the volunteered nature of OSM, the generated schema is ever-growing, flat, and heterogeneous. The tags do not convey semantics necessary for downstream applications. Aligning the tags to KG classes can help in incorporating rich semantics from knowledge graphs into OSM. We developed the NCA approach to align tags and classes by fusing the feature spaces of OSM and knowledge graphs. We utilized already linked entities to create a shared latent space that inherently captured the semantics of the tags. Furthermore, we probed the model to get the final tag-to-class alignment. We evaluated our approach for various country datasets of OSM, Wikidata, and DBpedia. Our experiments demonstrated that the NCA approach outperformed the state-of-the-art baselines by up to 13 and 37 percentage points on OSM-to-Wikidata and OSM-to-DBpedia tag-to-class alignment, respectively.

To summarize, our contributions are a novel approach to link class elements between OSM and knowledge graphs that utilizes a novel shared latent space that combines feature spaces of OSM and knowledge graphs. We also proposed a novel and effective algorithm to extract tag-to-class matches from the trained model.

7.1.2 Geographic Entity Alignment

In Chapter 5, we presented our approach for geographic entity alignment, where we aimed to align OSM objects to knowledge graph entities. Although OSM contains a huge amount of geographic entities, their representations are heterogeneous. The flat and sparse schema also hinders the applicability of OSM data in semantic applications. Knowledge graphs contain semantic information about geographic entities that can complement the OSM objects. Aligning OSM and knowledge graphs at the entity level provides knowledge graphs with precise geographic information and enables OSM objects with semantic information. However, schema and entity alignment are interrelated, wherein both enhance the performance of each other. In our IGEA approach, we build on the aforementioned principle of interrelation between schema and entity alignment. We developed a method to align schema and entity elements from OSM and knowledge graphs in an iterative manner. For the alignment, we utilized full entity descriptions as opposed to specific properties, that the state-of-the-art methods rely on. IGEA first applies the NCA approach to linked entities. The obtained tag-to-class alignments along with geographic distance are used to generate candidates for alignment pairs. A cross-attention-based classifier is applied to get the final matches. We then use the additional matches iteratively to improve on schema and entity alignment. We conducted our experiments on the OSM, Wikidata, and DBpedia knowledge graphs. The results of our experiments show that IGEA outperformed the baselines by up to 18 and 14 percentage points on Wikidata and DBpedia datasets in terms of F1-Score, respectively. We furthermore show an improvement of seven and eight percentage points in the results of tag-to-class alignment on Wikidata and DBpedia datasets as a result of using iterations.

In particular, our contributions are, a novel iterative cross-attention-based model to link the geographic entities between OSM and knowledge graphs. We also proposed a novel candidate blocking method that combines class-based and distance-based blocking.

7.1.3 Geographic Knowledge Graph Creation

In Chapter 6, we present a comprehensive knowledge graph created from OSM data. As mentioned earlier, in its original format, OSM data is not directly accessible to semantic applications. Furthermore, the flat OSM schema restricts the use of OSM in downstream applications. Having OSM data in the knowledge graph can mitigate the challenges that arise due to the flat OSM schema and the lack of semantics in the OMS tags. We convert the flat OSM schema into a novel hierarchical WorldKG ontology with superclass-subclass relationships. We then convert the OSM snapshot into a knowledge graph that conforms to the WorldKG ontology. The current version of WorldKG contains over 100 million entities belonging to over 33 super-classes and 1100 subclasses. WorldKG has over two magnitudes higher geographic entities when compared to the Wikidata and DBpedia knowledge graphs.

Our contributions are the WorldKG knowledge graph that lifts OSM data into semantic form which adheres to the WorldKG ontology. The novel WorldKG ontology represents OSM tags hierarchically and also presents links to the Wikidata and DBpedia class elements.

7.2 Future Outlook

In this thesis, we have developed novel approaches to make volunteered geographic data more accessible to downstream applications. Our research in geographic alignment and knowledge graph creation has led to following open research directions that can be further explored.

7.2.1 Enhancements to WorldKG

The current version of WorldKG contains mainly datatype properties, wherein the values at the object place are literals. As a result, the object-to-object links within the WorldKG knowledge graphs are not captured. For instance, for a triple *<wkg:Bonn wkgs:isin 'Germany'>*, Germany is represented as a literal value whereas it can be linked to the *wkg:Germany* entity. In the future, we would like to improve the connectivity within the WorldKG knowledge graph.

Another aspect that can improve the usability of WorldKG is to add more complex geometries, such as ways and relations from OSM to WorldKG. The current version of WorldKG only includes nodes of OSM, which are represented as point geometries. Complex geometries are represented using lines and polygons or multipolygons and can increase the complexity of the knowledge graph as well as the computation times of the services based on the knowledge graph, such as SPARQL endpoints. Including complex geographical objects such as relations and ways can enhance the expressiveness of WorldKG and can be further explored.

7.2.2 Development of Embedding Methods for Object Representation

Real-world applications such as accident prediction, and crime-rate prediction rely on region embeddings to get accurate results. These region embeddings are created using multiple data sources such as mobility data, land-usage data, and raster data. Using WorldKG data to build such region embeddings can be beneficial since WorldKG can provide interlinking with multiple sources such as Wikidata, and DBpedia. Using the spatial connections of WorldKG, these region embeddings can understand the important aspects of the place along with their descriptions.

7.2.3 Application to Geographic Question Answering

Geographic information retrieval applications such as location-based web search [AB07] or geographic question answering [Mai+20] have become popular recently. In the future, we can explore the integrated sources to enhance the answers given by geographic question answering systems.

Bibliography

[AB07] Dirk Ahlers and Susanne Boll. "Location-based Web Search". In: *The Geospatial Web, How Geobrowsers, Social Software and the Web 2.0 are Shaping the Network Society*. Springer, 2007, pp. 55–66. DOI: 10.1007/978-1-84628-827-2_6.

[Alg+19] Alsayed Algergawy, Daniel Faria, Alfio Ferrara, Irini Fundulaki, Ian Harrow, Sven Hertling, Ernesto Jiménez-Ruiz, Naouel Karam, Abderrahmane Khiat, Patrick Lambrix, Huanyu Li, Stefano Montanelli, Heiko Paulheim, Catia Pesquita, Tzanina Saveta, Pavel Shvaiko, Andrea Splendiani, Élodie Thiéblin, Cássia Trojahn, Jana Vatascinová, Ondrej Zamazal, and Lu Zhou. "Results of the Ontology Alignment Evaluation Initiative 2019". In: *Proceedings of the 14th International Workshop on Ontology Matching co-located with the 18th International Semantic Web Conference (ISWC 2019)*. Vol. 2536. 2019. URL: https://ceur-ws.org/Vol-2536/oaei19_paper0.pdf.

[ALH09] Sören Auer, Jens Lehmann, and Sebastian Hellmann. "LinkedGeoData: Adding a Spatial Dimension to the Web of Data". In: *Proceedings of the 8th International Semantic Web Conference, ISWC 2009*. Springer, 2009, pp. 731–746. DOI: 10.1007/978-3-642-04930-9_46.

[BG94] George Bebis and Michael Georgiopoulos. "Feed-forward neural networks". In: *IEEE Potentials* 13.4 (1994), pp. 27–31. DOI: 10.1109/45.329294.

[BK11] Robert Battle and Dave Kolas. "Geosparql: enabling a geospatial semantic web". In: *Semantic Web Journal* 3.4 (2011), pp. 355–370. URL: https://api.semanticscholar.org/CorpusID:31024156.

[Boj+17] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomás Mikolov. "Enriching Word Vectors with Subword Information". In: *Transactions of the Association for Computational Linguistics* 5 (2017), pp. 135–146. DOI: 10.1162/TACL_A_00051.

[BRR20] Ben Bondaruk, Steven A Roberts, and Colin Robertson. "Assessing the state of the art in Discrete Global Grid Systems: OGC criteria and present functionality". In: *Geomatica* 74.1 (2020), pp. 9–30. DOI: 10.1139/geomat-2019-0015.

[BZG20] Alexandre Bento, Amal Zouaq, and Michel Gagnon. "Ontology Matching Using Convolutional Neural Networks". In: *Proceedings of the 12th Language Resources and Evaluation Conference, LREC 2020*. European Language Resources Association, 2020, pp. 5648–5653. URL: https://aclanthology.org/2020.lrec-1.693/.

[Che+22] Wei Chen, Huaiyu Wan, Shengnan Guo, Haoyu Huang, Shaojie Zheng, Jiamu Li, Shuohao Lin, and Youfang Lin. "Building and exploiting spatial–temporal knowledge graph for next POI recommendation". In: *Knowledge-Based Systems* 258 (2022), p. 109951. ISSN: 0950-7051. DOI: 10.1016/j.knosys.2022.109951.

[CPT20] Riccardo Cappuzzo, Paolo Papotti, and Saravanan Thirumuruganathan. "Creating Embeddings of Heterogeneous Relational Datasets for Data Integration Tasks". In: *Proceedings of the 2020 International Conference on Management of Data, SIGMOD Conference 2020*. ACM, 2020, pp. 1335–1349. DOI: 10.1145/3318464.3389742.

[Der16] Leon Derczynski. "Complementarity, F-score, and NLP Evaluation". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016*. European Language Resources Association (ELRA), 2016. URL: http://www.lrec-conf.org/proceedings/lrec2016/summaries/105.html.

[Dev+19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019*. Association for Computational Linguistics, 2019, pp. 4171–4186. DOI: 10.18653/V1/N19-1423.

[Doa+04] AnHai Doan, Jayant Madhavan, Pedro M. Domingos, and Alon Y. Halevy. "Ontology Matching: A Machine Learning Approach". In: *Handbook on Ontologies*. International Handbooks on Information Systems. Springer, 2004, pp. 385–404. DOI: 10.1007/978-3-540-24750-0_19.

[Dso+21] Alishiba Dsouza, Nicolas Tempelmeier, Ran Yu, Simon Gottschalk, and Elena Demidova. "WorldKG: A World-Scale Geographic Knowledge Graph". In: *Proceedings of the 30th ACM International Conference on Information and Knowledge Management, CIKM 2021*. ACM, 2021, pp. 4475–4484. DOI: 10.1145/3459637.3482023.

[Dso+23] Alishiba Dsouza, Ran Yu, Moritz Windoffer, and Elena Demidova. "Iterative Geographic Entity Alignment with Cross-Attention". In: *Proceedings of the 22nd International Semantic Web Conference, ISWC 2023*. Springer, 2023, pp. 216–233. DOI: 10.1007/978-3-031-47240-4_12.

[DTD21] Alishiba Dsouza, Nicolas Tempelmeier, and Elena Demidova. "Towards Neural Schema Alignment for OpenStreetMap and Knowledge Graphs". In: *Proceeding of the 20th International Semantic Web Conference, ISWC 2021*. Springer, 2021, pp. 56–73. DOI: 10.1007/978-3-030-88361-4_4.

[Fer+13] Basura Fernando, Amaury Habrard, Marc Sebban, and Tinne Tuytelaars. "Unsupervised Visual Domain Adaptation Using Subspace Alignment". In: *IEEE International Conference on Computer Vision, ICCV 2013*. IEEE Computer Society, 2013, pp. 2960–2967. DOI: 10.1109/ICCV.2013.368.

[Fu+20] Cheng Fu, Xianpei Han, Jiaming He, and Le Sun. "Hierarchical Matching Network for Heterogeneous Entity Resolution". In: *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*. 2020, pp. 3665–3671. DOI: 10.24963/IJCAI.2020/507.

[Gan+16] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. "Domain-Adversarial Training of Neural Networks". In: *Journal of Machine Learning Research* 17 (2016), 59:1–59:35. DOI: 10.1007/978-3-319-58347-1_10.

[GDA12] GDAM. *Global Administrative Areas (2012)*. http://www.gadm.org/home. 2012.

[Goo07] Michael F. Goodchild. "Citizens as sensors: the world of volunteered geography". In: *GeoJournal* 69.4 (2007), pp. 211–221. DOI: 10.1007/s10708-007-9111-y.

[Gu+18] Jiuxiang Gu, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, Gang Wang, Jianfei Cai, and Tsuhan Chen. "Recent advances in convolutional neural networks". In: *Pattern Recognition* 77 (2018), pp. 354–377. DOI: 10.1016/J.PATCOG.2017.10.013.

[Gur97] Kevin N. Gurney. *An introduction to neural networks*. Morgan Kaufmann, 1997. ISBN: 978-1-85728-673-1. DOI: 10.1201/9781315273570.

[Hao+16] Yanchao Hao, Yuanzhe Zhang, Shizhu He, Kang Liu, and Jun Zhao. "A Joint Embedding Method for Entity Alignment of Knowledge Bases". In: *Proceedings of the Knowledge Graph and Semantic Computing: Semantic, Knowledge, and Linked Big Data - First China Conference, CCKS 2016*. Vol. 650. Springer, 2016, pp. 3–14. DOI: 10.1007/978-981-10-3168-7_1.

[HB11] Tom Heath and Christian Bizer. *Linked Data: Evolving the Web into a Global Data Space*. Synthesis Lectures on the Semantic Web. Morgan & Claypool Publishers, 2011. ISBN: 978-3-031-79431-5. DOI: 10.2200/S00334ED1V01Y201102WBE001.

[Hof+13] Johannes Hoffart, Fabian M. Suchanek, Klaus Berberich, and Gerhard Weikum. "YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia". In: *Artificial Intelligence* 194 (2013), pp. 28–61. DOI: 10.1016/J.ARTINT.2012.06.001.

[Jai+10] Prateek Jain, Pascal Hitzler, Amit P. Sheth, Kunal Verma, and Peter Z. Yeh. "Ontology Alignment for Linked Open Data". In: *Proceedings of the 9th International Semantic Web Conference, ISWC 2010*. Vol. 6496. Springer, 2010, pp. 402–417. DOI: 10.1007/978-3-642-17746-0_26.

[Jan+22] Krzysztof Janowicz, Pascal Hitzler, Wenwen Li, Dean Rehberger, Mark Schildhauer, Rui Zhu, Cogan Shimizu, Colby K. Fisher, Ling Cai, Gengchen Mai, Joseph Zalewski, Lu Zhou, Shirly Stephen, Seila Gonzalez Estrecha, Bryce D. Mecum, Anna Lopez-Carr, Andrew Schroeder, Dave Smith, Dawn J. Wright, Sizhe Wang, Yuanyuan Tian, Zilong Liu, Meilin Shi, Anthony D'Onofrio, Zhining Gu, and Kitty Currier. "Know, Know Where, Knowwheregraph: A Densely Connected, Cross-Domain Knowledge Graph and Geo-Enrichment Service Stack for Applications

in Environmental Intelligence". In: *AI Magazine* 43.1 (2022), pp. 30–39. DOI: 10.1609/AIMAG.V43I1.19120.

[KMK19] Nikolaos Karalis, Georgios M. Mandilaras, and Manolis Koubarakis. "Extending the YAGO2 Knowledge Graph with Precise Geospatial Knowledge". In: *Proceedings of the 18th International Semantic Web Conference, ISWC 2019*. Springer, 2019. DOI: 10.1007/978-3-030-30796-7_12.

[Kou+20] Christos Koutras, Marios Fragkoulis, Asterios Katsifodimos, and Christoph Lofi. "REMA: Graph Embeddings-based Relational Schema Matching". In: *Proceedings of the Workshops of the EDBT/ICDT 2020 Joint Conference, 2020*. Vol. 2578. CEUR-WS.org, 2020. URL: https://ceur-ws.org/Vol-2578/SEAData5.pdf.

[Lam+18] Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. "Word translation without parallel data". In: *6th International Conference on Learning Representations, ICLR 2018*, OpenReview.net, 2018. URL: https://openreview.net/forum?id=H196sainb.

[Leh+15] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. "DBpedia - A large-scale, multilingual knowledge base extracted from Wikipedia". In: *Semantic Web* 6.2 (2015), pp. 167–195. DOI: 10.3233/SW-140134.

[Li+09] Juanzi Li, Jie Tang, Yi Li, and Qiong Luo. "RiMOM: A Dynamic Multistrategy Ontology Alignment Framework". In: *IEEE Transactions on Knowledge and Data Engineering* 21.8 (2009), pp. 1218–1232. DOI: 10.1109/TKDE.2008.202.

[LLS16] Chun Lu, Philippe Laublet, and Milan Stankovic. "Travel Attractions Recommendation with Knowledge Graphs". In: *Proceedings of the 20th International Conference on Knowledge Engineering and Knowledge Management, EKAW 2016*. Vol. 10024. 2016, pp. 416–431. DOI: 10.1007/978-3-319-49004-5_27.

[Mai+20] Gengchen Mai, Krzysztof Janowicz, Ling Cai, Rui Zhu, Blake Regalia, Bo Yan, Meilin Shi, and Ni Lao. "SE-KGE : A location-aware Knowledge Graph Embedding model for Geographic Question Answering and Spatial Semantic Lifting". In: *Transactions of GIS* 24.3 (2020), pp. 623–655. DOI: 10.1111/TGIS.12629.

[MBR01] Jayant Madhavan, Philip A. Bernstein, and Erhard Rahm. "Generic Schema Matching with Cupid". In: *Proceedings of the 27th International Conference on Very Large Data Bases, VDLB 2001*. Morgan Kaufmann, 2001, pp. 49–58. URL: http://www.vldb.org/conf/2001/P049.pdf.

[MGR02] Sergey Melnik, Hector Garcia-Molina, and Erhard Rahm. "Similarity Flooding: A Versatile Graph Matching Algorithm and Its Application to Schema Matching". In: *Proceedings of the 18th International Conference on Data Engineering, ICDE 2002*. Ed. by Rakesh Agrawal and Klaus R. Dittrich. IEEE Computer Society, 2002, pp. 117–128. DOI: 10.1109/ICDE.2002.994702.

[Mik+13] Tomás Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. "Distributed Representations of Words and Phrases and their Compositionality". In: *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*. 2013, pp. 3111–3119.

[MRS08] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to information retrieval*. Cambridge University Press, 2008. ISBN: 978-0-521-86571-5. DOI: 10.1017/CBO9780511809071.

[Mud+18] Sidharth Mudgal, Han Li, theodoros Rekatsinas, AnHai Doan, Youngchoon Park, Ganesh Krishnan, Rohit Deep, Esteban Arcaute, and Vijay Raghavendra. "Deep Learning for Entity Matching: A Design Space Exploration". In: *Proceedings of the 2018 International Conference on Management of Data, SIGMOD 2018*. ACM, 2018, pp. 19–34. DOI: 10.1145/3183713.3196926.

[NA11] Axel-Cyrille Ngonga Ngomo and Sören Auer. "LIMES - A Time-Efficient Approach for Large-Scale Link Discovery on the Web of Data". In: *Proceedings of the 22nd International Joint Conference on Artificial Intelligence, 2011*. IJCAI/AAAI, 2011, pp. 2312–2317. DOI: 10.5591/978-1-57735-516-8/IJCAI11-385.

[NBT13] DuyHoa Ngo, Zohra Bellahsene, and Konstantin Todorov. "Opening the Black Box of Ontology Matching". In: *Proc. of the ESWC 2013*. 2013.

[Ngo13] Axel-Cyrille Ngonga Ngomo. "ORCHID - Reduction-Ratio-Optimal Computation of Geo-spatial Distances for Link Discovery". In: *Proceedings of the 12th International Semantic Web Conference, ISWC 2013*. Vol. 8218. Springer, 2013, pp. 395–410. DOI: 10.1007/978-3-642-41335-3_25.

[Nki+18] Ikechukwu Nkisi-Orji, Nirmalie Wiratunga, Stewart Massie, Kit-Ying Hui, and Rachel Heaven. "Ontology Alignment Based on Word Embedding and Random Forest Classification". In: *Proceedings of the Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2018*. Vol. 11051. Springer, 2018, pp. 557–572. DOI: 10.1007/978-3-030-10925-7_34.

[NL12] Axel-Cyrille Ngonga Ngomo and Klaus Lyko. "EAGLE: Efficient Active Learning of Link Specifications Using Genetic Programming". In: *The Semantic Web: Research and Applications - 9th Extended Semantic Web Conference, ESWC 2012*. Vol. 7295. Springer, 2012, pp. 149–163. DOI: 10.1007/978-3-642-30284-8_17.

[NL13] Axel-Cyrille Ngonga Ngomo and Klaus Lyko. "Unsupervised learning of link specifications: deterministic vs. non-deterministic". In: *Proceedings of the 8th International Workshop on Ontology Matching co-located with the 12th International Semantic Web Conference ISWC 2013*. Vol. 1111. CEUR-WS.org, 2013, pp. 25–36. URL: https://ceur-ws.org/Vol-1111/om2013_Tpaper3.pdf.

[NLC13] Axel-Cyrille Ngonga Ngomo, Klaus Lyko, and Victor Christen. "COALA - Correlation-Aware Active Learning of Link Specifications". In: *Proceedings of the 10th Extended Semantic Web Conference, ESWC, 2013*. Vol. 7882. Springer, 2013, pp. 442–456. DOI: 10.1007/978-3-642-38288-8_30.

[Noy+19] Natalya Fridman Noy, Yuqing Gao, Anshu Jain, Anant Narayanan, Alan Patterson, and Jamie Taylor. "Industry-scale knowledge graphs: lessons and challenges". In: *Communications of ACM* 62.8 (2019), pp. 36–43. DOI: 10.1145/3331166.

[NZ12] Pascal Neis and Alexander Zipf. "Analyzing the Contributor Activity of a Volunteered Geographic Information Project - The Case of OpenStreetMap". In: *ISPRS International Journal of Geo-Information* 1.2 (2012), pp. 146–165. DOI: 10.3390/IJGI1020146.

[NZ14] Pascal Neis and Dennis Zielstra. "Recent Developments and Future Trends in Volunteered Geographic Information Research: The Case of OpenStreetMap". In: *Future Internet* 6.1 (2014), pp. 76–106. DOI: 10.3390/FI6010076.

[ORG15] Lorena Otero-Cerdeira, Francisco Javier Rodríguez-Martínez, and Alma Gómez-Rodríguez. "Ontology matching: A literature review". In: *Expert Syst. Appl.* 42.2 (2015), pp. 949–971. DOI: 10.1016/J.ESWA.2014.08.032.

[Oun+21] Chahinez Ounoughi, Amira Mouakher, Muhammad Ibraheem Sherzad, and Sadok Ben Yahia. "A Scalable Knowledge Graph Embedding Model for Next Point-of-Interest Recommendation in Tallinn City". In: *Proceeding of the 15th International Conference on Research Challenges in Information Science RCIS 2021*. Vol. 415. Springer, 2021, pp. 435–451. DOI: 10.1007/978-3-030-75018-3_29.

[PB22] Ralph Peeters and Christian Bizer. "Supervised Contrastive Learning for Product Matching". In: *Companion Proceedings of the Web Conference, WWW 2022*. ACM, 2022, pp. 248–251. DOI: 10.1145/3487553.3524254.

[PMB13] Razvan Pascanu, Tomás Mikolov, and Yoshua Bengio. "On the difficulty of training recurrent neural networks". In: *Proceedings of the 30th International Conference on Machine Learning, ICML 2013*. Vol. 28. JMLR.org, 2013, pp. 1310–1318. URL: http://proceedings.mlr.press/v28/pascanu13.html.

[PSM14] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. "Glove: Global Vectors for Word Representation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*. ACL, 2014, pp. 1532–1543. DOI: 10.3115/V1/D14-1162.

[PT21] Nabin Paudyal and Arun Kumar Timalsina. "POI Recommendations with the Use of Knowledge Graph Convolutional Networks". In: (2021).

[Pun+18] Dharmen Punjani, Kuldeep Singh, Andreas Both, Manolis Koubarakis, Iosif Angelidis, Konstantina Bereta, Themis Beris, Dimitris Bilidas, Theofilos Ioannidis, Nikolaos Karalis, Christoph Lange, Despina-Athanasia Pantazi, Christos Papaloukas, and George Stamoulis. "Template-Based Question Answering over Linked Geospatial Data". In: *Proceedings of*

the 12th Workshop on Geographic Information Retrieval, GIR@SIGSPATIAL 2018. ACM, 2018, 7:1–7:10. DOI: 10.1145/3281354.3281362.

[Qiu+17] Lirong Qiu, Jia Yu, Qiumei Pu, and Chuncheng Xiang. "Knowledge entity learning and representation for ontology matching based on deep neural networks". In: *Cluster Computing* 20.2 (2017), pp. 969–977. DOI: 10.1007/S10586-017-0844-1.

[RB01] Erhard Rahm and Philip A. Bernstein. "A survey of approaches to automatic schema matching". In: *The International Journal on Very Large Data Bases VLDB* 10.4 (2001), pp. 334–350. DOI: 10.1007/S007780100057.

[RG19] Nils Reimers and Iryna Gurevych. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019*. Association for Computational Linguistics, 2019, pp. 3980–3990. DOI: 10.18653/V1/D19-1410.

[Sai+15] Tara N. Sainath, Oriol Vinyals, Andrew W. Senior, and Hasim Sak. "Convolutional, Long Short-Term Memory, fully connected Deep Neural Networks". In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015*. IEEE, 2015, pp. 4580–4584. DOI: 10.1109/ICASSP.2015.7178838.

[Sch+] Guus Schreiber, Frank Manola, Eric Miller, and Brian McBride. *RDF*. https://www.w3.org/TR/rdf11-primer/. [Online; accessed 10-March-2024].

[She+17] Mohamed Ahmed Sherif, Kevin Dreßler, Panayiotis Smeros, and Axel-Cyrille Ngonga Ngomo. "Radon - Rapid Discovery of Topological Relations". In: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, 2017*. AAAI Press, 2017, pp. 175–181. DOI: 10.1609/AAAI.V31I1.10478.

[SNL17] Mohamed Ahmed Sherif, Axel-Cyrille Ngonga Ngomo, and Jens Lehmann. "Wombat - A Generalization Approach for Automatic Link Discovery". In: *Proceedings of the Semantic Web - 14th International Conference, ESWC 2017*. Vol. 10249. Springer, 2017, pp. 103–119. DOI: 10.1007/978-3-319-58068-5_7.

[Ste+12] Thomas Steiner, Ruben Verborgh, Raphaël Troncy, Joaquim Gabarró, and Rik Van de Walle. "Adding Realtime Coverage to the Google Knowledge Graph". In: *Proceedings of the ISWC 2012 Posters & Demonstrations Track, 2012*. Vol. 914. CEUR-WS.org, 2012. URL: https://ceur-ws.org/Vol-914/paper_2.pdf.

[TD21] Nicolas Tempelmeier and Elena Demidova. "Linking OpenStreetMap with knowledge graphs - Link discovery for schema-agnostic volunteered geographic information". In: *Future Generation Computer Systems* 116 (2021), pp. 349–364. DOI: 10.1016/J.FUTURE.2020.11.003.

[Vas+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. "Attention is All you Need". In: *Proceedings of the Annual Conference on Neural Information Processing Systems, 2017*. 2017, pp. 5998–6008.

URL: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[VK14] Denny Vrandecic and Markus Krötzsch. "Wikidata: a free collaborative knowledgebase". In: *Communications of the ACM* 57.10 (2014), pp. 78–85. DOI: 10.1145/2629489.

[Vol+09] Julius Volz, Christian Bizer, Martin Gaedke, and Georgi Kobilarov. "Silk - A Link Discovery Framework for the Web of Data". In: *Proceedings of the WWW2009 Workshop on Linked Data on the Web, LDOW 2009*. Vol. 538. CEUR-WS.org, 2009. URL: https://ceur-ws.org/Vol-538/ldow2009_paper13.pdf.

[Wei+20] Xi Wei, Tianzhu Zhang, Yan Li, Yongdong Zhang, and Feng Wu. "Multi-Modality Cross Attention Network for Image and Sentence Matching". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020*. IEEE, 2020, pp. 10938–10947. DOI: 10.1109/CVPR42600.2020.01095.

[Xia+15] Chuncheng Xiang, Tingsong Jiang, Baobao Chang, and Zhifang Sui. "ERSOM: A Structural Ontology Matching Approach Using Automatically Learned Entity Representation". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015*. The Association for Computational Linguistics, 2015, pp. 2419–2429. DOI: 10.18653/V1/D15-1289.

[YJY24] Jonghyeon Yang, Hanme Jang, and Kiyun Yu. "Geographic Knowledge Base Question Answering over OpenStreetMap". In: *ISPRS International Journal of Geo-Information* 13.1 (2024). ISSN: 2220-9964. DOI: 10.3390/ijgi13010010.

Appendices

Appendix A

Publication: Towards Neural Schema Alignment for OpenStreetMap and Knowledge Graphs

Towards Neural Schema Alignment for OpenStreetMap and Knowledge Graphs

Alishiba Dsouza[1( )], Nicolas Tempelmeier[2], and Elena Demidova[1]

[1] Data Science and Intelligent Systems (DSIS), University of Bonn, Bonn, Germany
{dsouza,elena.demidova}@cs.uni-bonn.de
[2] L3S Research Center, Leibniz Universität Hannover, Hannover, Germany
tempelmeier@L3S.de

Abstract. OpenStreetMap (OSM) is one of the richest, openly available sources of volunteered geographic information. Although OSM includes various geographical entities, their descriptions are highly heterogeneous, incomplete, and do not follow any well-defined ontology. Knowledge graphs can potentially provide valuable semantic information to enrich OSM entities. However, interlinking OSM entities with knowledge graphs is inherently difficult due to the large, heterogeneous, ambiguous, and flat OSM schema and the annotation sparsity. This paper tackles the alignment of OSM tags with the corresponding knowledge graph classes holistically by jointly considering the schema and instance layers. We propose a novel neural architecture that capitalizes upon a shared latent space for tag-to-class alignment created using linked entities in OSM and knowledge graphs. Our experiments aligning OSM datasets for several countries with two of the most prominent openly available knowledge graphs, namely, Wikidata and DBpedia, demonstrate that the proposed approach outperforms the state-of-the-art schema alignment baselines by up to 37% points F1-score. The resulting alignment facilitates new semantic annotations for over 10 million OSM entities worldwide, which is over a 400% increase compared to the existing annotations.

Keywords: OpenStreetMap · Knowledge graph · Neural schema alignment

1 Introduction

OpenStreetMap (OSM) has evolved as a critical source of openly available geographic information globally, including rich data from 188 countries. This information is contributed by a large community, currently counting over 1.5 million volunteers. OSM captures a vast and continuously growing number of geographic entities, currently counting more than 6.8 billion [15]. The descriptions of OSM entities consist of heterogeneous key-value pairs, so-called *tags*, and include over 80 thousand distinct keys. OSM keys and tags do not possess machine-readable semantics, such that OSM data is not directly accessible for semantic applications. Whereas knowledge graphs (KGs) can provide precise semantics for geographic entities, large publicly available general-purpose KGs like Wikidata [30],

A. Hotho et al. (Eds.): ISWC 2021, LNCS 12922, pp. 56–73, 2021.
https://doi.org/10.1007/978-3-030-88361-4_4

DBpedia [2], YAGO [26], and specialized KGs like EventKG [10], and LinkedGeoData [25] lack coverage of geographic entities. For instance, in June 2021, 931,574 entities with tag `amenity=restaurant` were present in OSM, whereas Wikidata included only 4,391 entities for the equivalent class "restaurant".

An alignment of OSM and knowledge graphs at the schema level can make a wide variety of geographic entities in OSM accessible through semantic technologies and applications. The automatic suggestions of alignment candidates can help to create accurate schema mappings in human-in-the-loop applications. Furthermore, alignment models can help OSM volunteers to map geographic entities in OSM and annotate these entities with KG classes.

The problem of schema alignment between OSM and KGs is particularly challenging due to several factors, most prominently including the heterogeneous representations of types and properties of geographic entities via OSM tags, unclear tag semantics, the large scale and flatness of the OSM schema, and the sparseness of the existing links. OSM does not limit the usage of keys and tags by any strict schema and provides only a set of guidelines[1]. As a result, the types and properties of OSM entities are represented via a variety of tags that do not possess precise semantics. Consider an excerpt from the representations of the entity "Zugspitze" (mountain in Germany) in Wikidata and OSM:

Wikidata

Subject	**Predicate**	**Object**
*Q*3375	*label*	*Zugspitze*
*Q*3375	*coordinate*	$47°25'N, 10°59'E$
*Q*3375	*parentpeak*	*Q*15127
*Q*3375	*instance of*	*mountain*

OpenStreetMap

Key	**Value**
id	27384190
`name`	*Zugspitze*
`natural`	*peak*
`summit:cross`	*yes*

In Wikidata, an entity type is typically represented using the `instance of` property. In this example, the statement "Q3375 `instance of` mountain" indicates the type "mountain" of the entity "Q3375". In OpenStreetMap, the type "mountain" of the same entity is indicated by the tag `natural=peak`. As OSM lacks a counterpart of the `instance of` property, it is unclear which particular tag represents an entity type and which tags refer to other properties. Furthermore, multiple OSM tags can refer to the same semantic concept. Finally, whereas the OSM schema with over 80 thousand distinct keys is extensive, the alignment between OSM and knowledge graphs at the schema level is almost nonexistent. For instance, as of April 2021, Wikidata contained 585 alignments between its properties and OSM keys, corresponding to only 0.7% of the distinct OSM keys. Overall, the flatness, heterogeneity, ambiguity, and the large scale of OSM schema, along with a lack of links, make the alignment particularly challenging.

Existing approaches for schema alignment operate at the schema and instance level and consider the similarity of schema elements, structural similarity, and instance similarity. As OSM schema is flat, ontology alignment methods that utilize hierarchical structures, such as [13,17], are not applicable. A transformation of OSM data into a tabular or relational format leads to highly sparse tables with

[1] OSM "How to map a": https://wiki.openstreetmap.org/wiki/How_to_map_a.

numerous columns. Therefore, approaches to syntactic or instance-based alignment for relational or tabular data, such as e.g., [6,32], or syntactic matching of schema element names [28] cannot yield good results for matching OSM tags with KG classes.

This paper takes the first important step to align OSM and knowledge graphs at the schema level using a novel neural method. In particular, we tackle tag-to-class alignment, i.e., we aim to identify OSM tags that convey class information and map them to the corresponding classes in the Wikidata knowledge graph and the DBpedia ontology. We present the Neural Class Alignment (NCA) model - a novel instance-based neural approach that aligns OSM tags with the corresponding semantic classes in a knowledge graph. NCA builds upon a novel shared latent space that aligns OSM tags and KG concepts and facilitates a seamless translation between them. To the best of our knowledge, NCA is the first approach to align OSM and KGs at the schema level with a neural method.

Our contributions are as follows:

- We present NCA – a novel approach for class alignment for OSM and KGs.
- We propose a novel shared latent space that fuses feature spaces from knowledge graphs and OSM in a joint model, enabling simultaneous training of the schema alignment model on heterogeneous semantic and geographic sources.
- We develop a novel, effective algorithm to extract tag-to-class alignments from the resulting model.
- The results of our evaluation demonstrate that the proposed NCA approach is highly effective and outperforms the baselines by up to 37% points F1-score.
- As a result of the proposed NCA alignment method, we provide semantic annotations with Wikidata and DBpedia classes for over 10 million OSM entities. This result corresponds to an over 400% increase compared to currently existing annotations.
- We make our code and datasets publicly available and provide a manually annotated ground truth for the tag-to-class alignment of OSM tags with Wikidata and DBpedia classes[2].

2 Problem Statement

In this section, we formalize the problem definition. First, we formally define the concepts of an OSM corpus and a knowledge graph. An OSM corpus contains nodes representing geographic entities. Each node is annotated with an identifier, a location, and a set of key-value pairs known as tags.

Definition 1. *An* OSM corpus $\mathcal{C} = (N, T)$ *consists of a set of nodes* N *representing geographic entities, and a set of tags* T*. Each tag* $t \in T$ *is represented as a key-value pair, with the key* $k \in K$ *and a value* $v \in V$*:* $t = \langle k, v \rangle$*. A node* $n \in N$*,* $n = \langle i, l, T_n \rangle$ *is represented as a tuple containing an identifier* i*, a geographic location* l*, and a set of tags* $T_n \subset T$*.*

[2] GitHub repository: https://github.com/alishiba14/NCA-OSM-to-KGs.

A knowledge graph contains real-world entities, classes, properties, and relations.

Definition 2. *A* knowledge graph $\mathcal{KG} = (E, C, P, L, F)$ *consists of a set of entities* E*, a set of classes* $C \subset E$*, a set of properties* P*, a set of literals* L*, and a set of triples* $F \subseteq E \times P \times (E \cup L)$*.*

The entities in E represent real-world entities and semantic classes. The properties in P represent relations connecting two entities, or an entity and a literal value. An entity in a KG can belong to one or multiple classes. An entity is typically linked to its class using the `rdf:type`, or an equivalent property.

Definition 3. *A class of the entity* $e \in E$ *in the knowledge graph* $\mathcal{KG} = (E, C, P, L, F)$ *is denoted as:* $class(e) = \{c \in C \mid (e, \texttt{rdf:type}, c) \in F\}$*.*

An OSM node and a KG entity referring to the same real-world geographic entity and connected via an identity link are denoted linked entities.

Definition 4. *A linked entity* $(n, e) \in E_L$ *is a pair of an OSM node* $n = \langle i, l, T_n \rangle$*,* $n \in N$*, and a knowledge graph entity* $e \in E$ *that corresponds to the same real-world entity. In a knowledge graph, a linked entity is typically represented using a* $(e, \texttt{owl:sameAs}, i)$ *triple, where* i *is the node identifier.* E_L *denotes the set of all linked entities in a knowledge graph.*

This paper tackles the alignment of tags that describe node types in an OSM corpus to equivalent classes in a knowledge graph.

Definition 5. ***Tag-to-class alignment:*** *Given a knowledge graph* $\mathcal{KG}$ *and an OSM corpus* $\mathcal{C}$*, find a set of pairs* $tag\_class \subseteq (T \times C)$ *of OSM tags* T *and the corresponding* $\mathcal{KG}$ *classes, such that for each pair* $(t, c) \in tag\_class$ *OSM nodes with the tag* t *belong to the class* c*.*

3 Neural Class Alignment Approach

An alignment of an OSM corpus with a knowledge graph can include several dimensions, such as entity linking, node classification (i.e., aligning OSM nodes with the corresponding semantic classes in a knowledge graph), as well as alignment of schema elements such as keys/tags and the corresponding semantic classes. The alignments in these dimensions can reinforce each other. For example, linking OSM nodes with knowledge graph entities and classifying OSM nodes into knowledge graph classes can lead to new schema-level alignments and vice versa. Our proposed NCA approach systematically exploits the existing identity links between OSM nodes and knowledge graph entities based on this intuition. NCA builds an auxiliary classification model and utilizes this model to align OSM tags with the corresponding classes in a knowledge graph ontology.

NCA is an unsupervised two-step approach for tag-to-class alignment. Figure 1 presents an overview of the proposed NCA architecture. First, we build

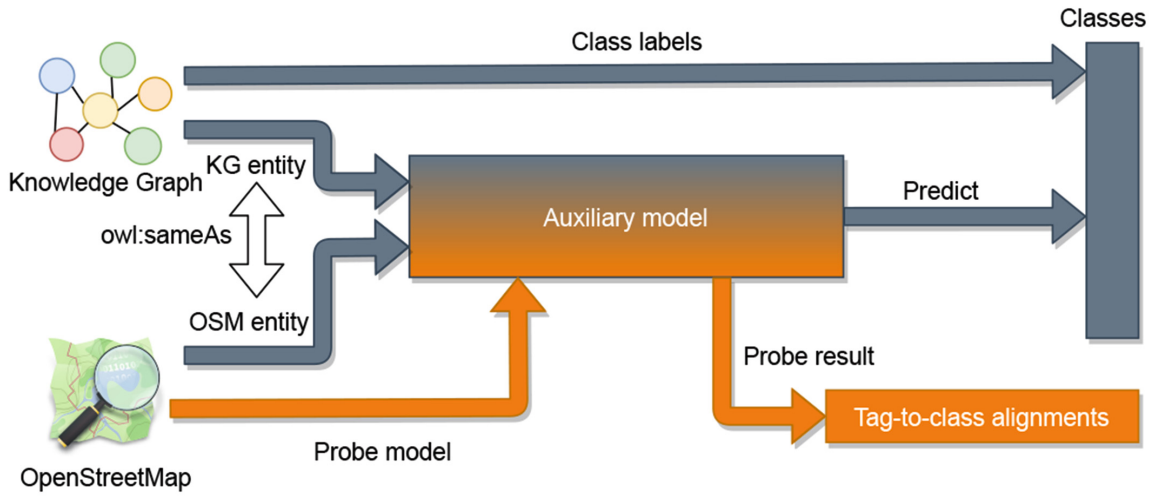


Fig. 1. Overview of the NCA architecture. The gray color indicates the first step (training of the auxiliary classification model). The orange color indicates the second step, i.e., the extraction of tag-to-class alignments. (Color figure online)

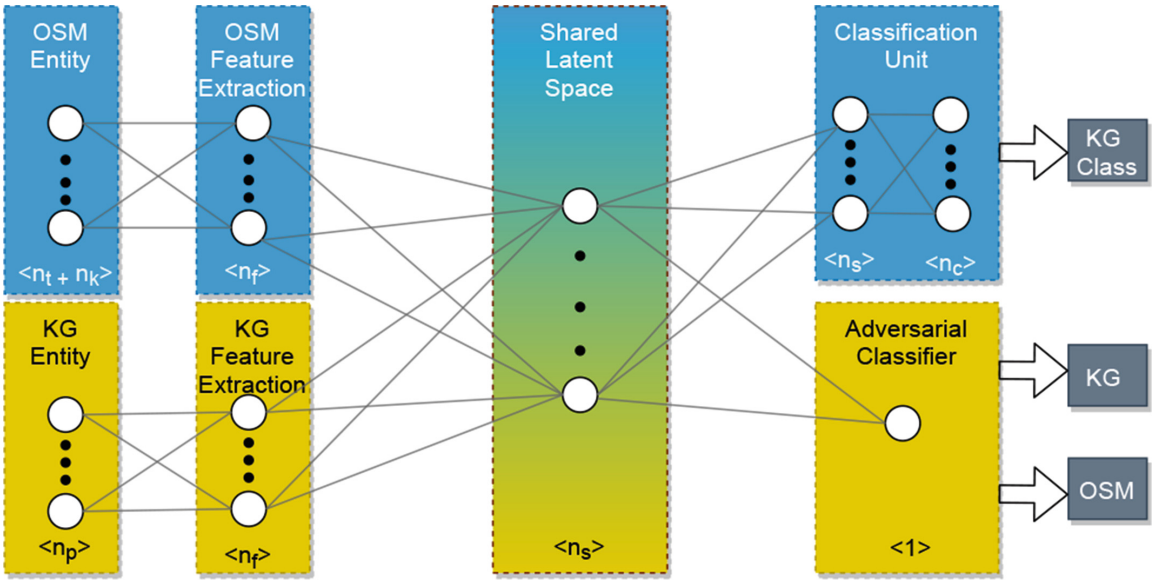


Fig. 2. The auxiliary classification model architecture. The blue color indicates the KG classification component, yellow marks the adversarial entity discrimination component. Parameters inside angular brackets denote the number of neurons in each layer, and lines denote the fully connected layers. (Color figure online)

an auxiliary neural classification model and train this model using linked entities in OSM and a KG. As a result, the model learns a novel shared latent space that aligns the feature spaces of OSM and a knowledge graph and implicitly captures tag-to-class alignments. Second, we systematically probe the resulting model to identify the captured alignments.

3.1 Auxiliary Neural Classification Model

In this step, we build a supervised auxiliary neural classification model for a dummy task of OSM node and KG entity classification. The model resulting from this step is later used for the tag-to-class alignment. Figure 2 presents the model architecture. The parameters n_t, n_k, n_p, n_c denote the number of OSM tags, number of OSM keys, number of KG properties, and number of KG classes, respectively. We experimentally select the number of neurons in the feature extraction layer (n_f) and the shared latent space layer (n_s). The auxiliary classification model architecture consists of several components described below.

OSM Node Representation. We represent an OSM node as a binary vector in an **O**-dimensional vector space. The space dimensions correspond to OSM tags or keys, and binary values represent whether the node includes the corresponding tag or key. The vector space dimensions serve as features for the classification model, such that we also refer to this space as the OSM feature space. To select the most descriptive tags to be included as dimensions in the OSM feature space, we filter out low-quality tags using OSM taginfo[3]. We include only the tags with an available description in the OSM wiki[4] having at least 50 occurrences within OSM. For tags with infrequent values (e.g., literals), we include only the keys as dimensions. We aim to align geographic concepts and not specific entities; thus, we do not include infrequent and node-specific values such as entity names or geographic coordinates in the representation. For instance, the concept of "mountain" is the same across different geographic regions, such that the geographic location of entities is not informative for the schema alignment.

KG Entity Representation. We represent a KG entity as a binary vector in a **V**-dimensional vector space. The space dimensions correspond to the KG properties. Binary values represent whether the entity includes the corresponding property. The vector space dimensions serve as features for the classification model, such that we also refer to this space as the KG feature space. To select the most descriptive properties to be included in the KG feature space, we rank the properties based on their selectivity concerning the class and the frequency of property usage (i.e., the number of statements in the KG that assign this property to an entity). Given a property p, we calculate its weight as: $weight(p, c) = n_{p,c} * log\frac{N}{c_p}$. Here, $n_{p,c}$ denotes the number of statements in which the property p is assigned to an entity of class c, N denotes the total number of classes in a knowledge graph, and c_p is the number of distinct classes that include the property p. For each class c, we select top-25 properties as features. These properties are included as dimensions in the KG feature space.

OSM & KG Feature Extraction. The KG and OSM feature representations serve as input to the specific fully connected feature extraction layers: OSM feature extraction and KG feature extraction. The purpose of these layers is to refine the vector representations obtained in the previous step.

[3] OSM taginfo: https://taginfo.openstreetmap.org/tags.
[4] OSM wiki: https://wiki.openstreetmap.org/wiki/.

Shared Latent Space & Adversarial Classifier. We introduce a novel *shared latent space* that fuses the initially disjoint feature spaces of OSM and KG such that entities from both data sources are represented in a joint space similarly. In addition to the training on OSM examples, shared latent space enables us to train our model on the KG examples. These examples provide the properties known to indicate class information [21]. The shared latent space component consists of a fully connected layer that receives the input from the OSM and KG feature extraction layers. Following recent domain adaption techniques [9], we use an adversarial classification layer to align latent representations of KG and OSM entities. The objective of the adversarial classifier is to discriminate whether the current training example is an OSM node or a KG entity, where the classification loss is measured as binary cross-entropy.

$$BinaryCrossEntropy = -\frac{1}{n}\sum_{i=1}^{n}[y_i \times \log(\hat{y}_i) + (1 - y_i) \times \log(1 - \hat{y}_i)],$$

where n is the total number of examples, y_i is the true class label, and $\hat{y}_i$ is the predicted class label. Intuitively, in a shared latent space, the classifier should not be able to distinguish whether a training example originates from OSM or a KG. To fuse the initially disjoint feature spaces, we reverse the gradients from the adversarial classification loss: $\mathcal{L}_{adverse} = -BinaryCrossEntropy_{adverse}$.

Classification Unit. To train the auxiliary classification model for the OSM nodes, we exploit linked entities. We label OSM nodes with semantic classes of equivalent KG entities. We use these class labels as supervision in the OSM node classification task. More formally, given a linked entity, $(n, e) \in E_L$, the training objective of the model is to predict $class(e)$ from n. Analogously, the training objective for a KG entity e is to predict the class label $class(e)$ of this entity.

We utilize a 2-layer feed-forward network as a classification model. In the last prediction layer of this network, each neuron corresponds to a class. As an entity can be assigned to multiple classes, we use a sigmoid activation function and a binary cross-entropy loss to achieve multi-label classification: $\mathcal{L}_{classifcation} = BinaryCrossEntropy_{classification}$. Finally, the joint loss function $\mathcal{L}$ of the network is given by $\mathcal{L} = \mathcal{L}_{classifcation} + \mathcal{L}_{adverse}$. In the training process, we alternate OSM and KG instances to avoid bias towards one data source.

3.2 Tag-to-Class Alignment

In this step, we systematically probe the trained auxiliary classification model to extract the tag-to-class alignment. The goal of this step is to obtain the corresponding KG class for a given OSM tag. Algorithm 1 details the extraction process. First, we load the pre-trained auxiliary model m (line 1) and initialize the result set (line 2). We then probe the model with a given list of OSM tags $\mathcal{T}$ (line 3). For a single tag $t \in \mathcal{T}$, we feed t to the OSM input layer of the auxiliary

Algorithm 1. Extract Tag-to-Class Alignment

`Input:`	m	Trained auxiliary model
	$\mathcal{T}$	List of OSM tags
	th_a	Alignment threshold
`Output:`	$align \subseteq (T \times C)$	Extracted alignment of tags and classes

1: `load`(m)
2: $align \Leftarrow \emptyset$
3: **for all** $t \in \mathcal{T}$ **do**
4: `forward_propagation(t, m)`
5: $activations \Leftarrow$ `extract_activations`(m)
6: **for all** $a \in activations$ **do**
7: **if** $a > th_a$ **then**
8: $align \Leftarrow align \cup \{(t, \texttt{class}(a))\}$
9: **end if**
10: **end for**
11: **end for**
12: **return** $align$

model and compute the complete forward propagation of t within m (line 4). We then extract the activation of the neurons of the last layer of the classification model before the sigmoid nonlinearity (line 5). As the individual neurons in this layer directly correspond to KG classes, we expect that the activation of the specific neurons quantifies the likeliness that the tag t corresponds to the respective class. For each activation of a specific neuron a that is above the alignment threshold th_a (line 6–7), we extract the corresponding class c and add this class to the set of alignments (line 8). We determine the threshold value experimentally, as described later in Sect. 5.3. As an OSM tag can have multiple corresponding classes, we opt for all matches above the threshold value. Finally, the resulting set *align* constitutes the inferred tag-to-class alignments.

3.3 Illustrative Example

We illustrate the proposed NCA approach at the example of the "Zugspitze" mountain introduced in Sect. 1. We create the representation of the Wikidata object "Q3375" in the KG feature space by creating a binary vector that has ones in the dimensions that correspond to the properties that this entity contains, such as, `label, coordinate, parentpeak`, and zeros otherwise. Note that the `instance of` predicate is not included in the feature space, as this predicate represents the class label. Similarly, we encode the OSM node with the id "27384190" in the OSM feature space by creating a vector that includes `name, natural=peak, summit:cross` as ones, and zeros in all other dimensions. As described above, we use frequent key-value pairs such as `natural=peak` as features, whereas for the infrequent key-value pairs, such as `name=Zugspitze`, we use only the key (i.e., `name`) as a feature. The KG and OSM features spaces

are then aligned in the shared latent space. To form this space, we train the auxiliary classification model that learns to output the correct class labels, such as "mountain". In the last prediction layer of this model, each neuron corresponds to a class. After the training is completed, we probe the classification model with a single tag, such as `natural=peak`. The activation of the neurons in the prediction layer corresponds to the predicted tag-to-class mapping. We output all classes with the activation values above the threshold th_a (here: "mountain").

4 Evaluation Setup

This section introduces the evaluation setup regarding datasets, ground truth generation, baselines, and evaluation metrics. All experiments were conducted on an AMD Opteron 8439 SE processor @ 2.7 GHz and 252 GB of memory, whereas the execution of NCA required up to 16 GB of memory only.

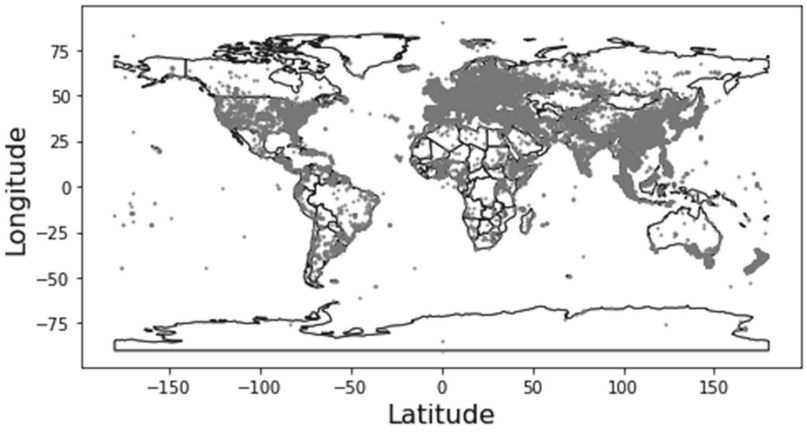


Fig. 3. OSM and Wikidata linked entities located on a world map.

4.1 Datasets

We carry out our experiments on OSM, Wikidata [30], and DBpedia [2] datasets.

Knowledge Graphs: A sufficient number of linked entities and distinct classes is essential to train the proposed neural model and achieve a meaningful schema alignment. As illustrated in Fig. 3, OSM to Wikidata links are highly frequent in the European region. We systematically rank European countries according to the number of linked entities between OSM and knowledge graphs. We choose the top-4 countries having at least ten distinct classes in the linked entity set. Based on these criteria, we select the Wikidata datasets for France, Germany, Great Britain, and Russia as well as the DBpedia datasets for France, Germany, Great Britain, and Spain. Although over 100,000 entity links between Russian DBpedia and OSM exist, most entities belong to only two classes. Hence, we omit Russian DBpedia from our analysis. Additionally, to understand the effect of NCA in other parts of the world, we select the USA and Australia with a

moderate amount of KG links. In our experiments, we consider Wikidata and DBpedia snapshots from March 2021. We collect the data from knowledge graphs by querying their SPARQL endpoints. We only consider geographic entities, i.e., the entities with valid geographic coordinates.

OpenStreetMap: We extract OSM data for France, Germany, Great Britain, Spain, Russia, the USA, and Australia. To facilitate evaluation, we only consider OSM nodes which include links to knowledge graphs. The number of entities assigned to specific knowledge graph classes follows a power-law distribution. We select the classes with more than 100 entities (i.e., 3% of classes in Wikidata) to facilitate model training. Note that some KG entities are linked to more than one OSM node, such that the number of nodes and entities in the dataset differ.

4.2 Ground Truth Creation

For Wikidata, we start the creation of our ground truth based on the "OpenStreetMap tag or key" Wikidata property[5]. This property provides a link between a Wikidata class and the corresponding OSM tag. However, this dataset is incomplete and lacks some language-specific classes as well as superclass and subclass relationships based on our manual analysis. We manually extended the ground truth by checking all possible matches obtained by the proposed NCA approach and all baseline models used in the evaluation. We added all correct matches to our ground truth. For DBpedia, we constructed the ground truth manually by labeling all combinations ($T \times C$) of OSM tags t and $\mathcal{KG}$ classes C in our dataset. For both KGs, we consider region-specific matches ("Ortsteil" vs. "District") and subclass/superclass relations (e.g., "locality" vs. "city/village").

4.3 Baselines

The schema alignment task of OSM and KG has not been addressed before, such that no task-specific baseline exists. For evaluation, we choose the state-of-the-art baselines from schema alignment for tabular data (Cupid [13], EmbDI [5], Similarity Flooding [14]), which is the closest representation to the OSM flat schema structure. Furthermore, we evaluate string similarity using Levenshtein distance, word embeddings-based cosine similarity, and SD-Type [21] - an established approach for type inference. To fit our data to the baselines, we convert our OSM (source) data and KG (target) data into a tabular format. For OSM, we use the tags and keys as columns and convert each node into a row. Similarly, for KGs, the properties and classes are converted into columns, and the entities form the rows. We evaluate our proposed method against the following baselines:

Cupid: Cupid [13] matches schema elements based on element names, structure, and data types. Cupid is a 2-phase approach. The first phase calculates the lexicographic similarity of names and data types. The second phase matches

[5] Wikidata "OpenStreetMap tag or key" property: https://www.wikidata.org/wiki/Property:P1282.

elements using the structural similarity based on the element proximity in the ontology hierarchy. As the OSM schema is flat, we consider a flat hierarchy, where the OSM table is the root and all columns are child nodes. The final Cupid score is the average similarity between the two phases.

Levenshtein Distance (LD): The Levenshtein distance (edit distance) is a string-based similarity measure used to match ontology elements lexicographically. The Levenshtein distance between two element names is calculated as the minimal number of edits needed to transform one element name to obtain the other. The modifications include addition, deletion, or replacement of characters [28]. We calculate the Levenshtein distance between all pairs of class names and tags and accept pairs with a distance lower than the threshold $th_l \in [0, 1]$.

EmbDi: EmbDi [5] is an algorithm for schema alignment and entity resolution. The algorithm maps table rows to a directed graph based on rows, columns, and cell values. EmbDi infers column embeddings by performing random walks on the graph. The random walks form sentences that constitute an input to a Word2Vec model. Finally, the similarity of the two columns is measured as the cosine similarity of the respective embeddings.

Similarity Flooding (SF): Similarity Flooding [14] transforms a data table into a directed labeled graph in which the nodes represent table columns. The weights of graph edges represent the node similarity, initialized using string similarity of the column names. The algorithm refines the weights by iteratively propagating similarity values along the edges. Each pair of nodes connected with a similarity value above the matching threshold forms an alignment.

SD-Type (SD): SD-Type [21] is an established approach for type inference. While SD-type was originally proposed to infer instance types based on conditional probabilities, we transfer the idea to infer class types. We calculate the conditional probability of a tag t given a class c as follows: $p(c|t) = \frac{\sum(t \bigcap c)}{\sum t}$. We accept all the matches with the probability values above threshold $th_l \in [0, 1]$.

Word Embedding Based Cosine Similarity (WECS): We use pre-trained word embeddings[6] trained using fastText [4] with 300 dimensions to obtain the word vectors of tag and class names. We calculate the cosine similarity between the word vectors of each tag-class pair. We accept all pairs with cosine similarity above the threshold $th_l \in [0, 1]$ as a match.

For LD, SD and WECS, we apply an exhaustive grid search to optimize the value of th_l for each dataset and report the highest resulting F1-scores. For the Cupid, EmbDi, and SF baseline implementation, we use the source code from the delftdata GitHub repository[7].

4.4 Metrics

The standard evaluation metrics for schema alignment are precision, recall, and F1-score computed against a reference alignment (i.e., ground truth). We eval-

[6] https://dl.fbaipublicfiles.com/fasttext/vectors-crawl/cc.en.300.bin.gz.
[7] Delftdata GitHub repository: https://github.com/delftdata/valentine.

uate the mappings as pairs, where each pair consists of one tag and one class (tag-to-class alignment). **Precision** is the fraction of correctly identified pairs among all identified pairs. **Recall** is the fraction of correctly identified pairs among all pairs in the reference alignment. **F1-score** is the harmonic mean of recall and precision. We consider the F1-score to be the most relevant metric since it reflects both precision and recall.

5 Evaluation

The evaluation aims to assess the performance of the proposed NCA approach for tag-to-class alignment in terms of precision, recall, and F1-score. Furthermore, we aim to analyze the influence of the confidence threshold and the impact of the shared latent space on the alignment performance. Note that we do not evaluate the artificial auxiliary classification task. Instead, we evaluate the utility of the auxiliary model in the overall schema alignment task. We train and evaluate the models for each country and knowledge graph separately.

5.1 Tag-to-Class Alignment Performance

Table 1 and 2 summarize the performance results of the baselines and our proposed NCA approach with respect to precision, recall and F1-score for tag-to-class alignment of OSM tags to Wikidata and DBpedia classes, respectively. As we can observe, the proposed NCA approach outperforms the baselines in terms of F1-score on all datasets. On Wikidata, we achieve up to 13% points F1-score improvement and ten percentage points on average compared to the best baseline. On DBpedia, we achieve up to 37% points F1-score improvement and 21% points on average. As OSM lacks a hierarchical structure, limiting structural comparison, most of the applicable baselines build on the name comparison. Here, the heterogeneity of OSM tags limits the precision of the baselines substantially. SD-Type obtains the highest F1-score amongst baselines. NCA uses the property, tags, and keys information from the shared latent space and achieves higher performance than the best performing SD-Type baseline. For other baselines, the absolute values achieved are relatively low. SF, WECS, and EmbDI obtain only low similarity values, resulting in low precision. An increase of the confidence threshold for these baselines leads to zero matches. The tag-class pairs vary significantly in terms of linguistic and semantic similarities. The correct pairs obtained using WECS do not obtain sufficiently high scores to discriminate from the wrong matches, making WECS one of the weakest baselines.

We observe performance variations across countries and knowledge graphs, with Australian Wikidata and French DBpedia achieving the highest F1-scores compared to other countries. These variations can be explained by the differences in the dataset characteristics, including the number of links, entities per class, and unique tags and classes per country. These characteristics vary significantly across the datasets. Furthermore, the number of classes per entity varies. On average, Wikidata indicates one class per entity (i.e., the most specific class).

Table 1. Tag-to-class alignment performance for OSM tags to Wikidata classes.

Name	France			Germany			Great Britain			Russia			USA			Australia			Average		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
CUPID	0.06	1.00	0.12	0.03	0.70	0.06	0.07	1.00	0.14	0.08	0.80	0.15	0.06	1.00	0.11	0.25	1.00	0.38	0.09	0.91	0.16
LD	0.45	0.28	0.35	0.65	0.34	0.44	0.54	0.37	0.44	0.64	0.34	0.45	0.39	0.37	0.38	0.31	0.41	0.36	0.49	0.35	0.40
EMBDI	0.03	1.00	0.06	0.02	1.00	0.03	0.04	1.00	0.06	0.02	1.00	0.03	0.01	1.00	0.03	0.08	0.91	0.15	0.05	0.98	0.06
SF	0.03	1.00	0.06	0.02	1.00	0.03	0.01	1.00	0.03	0.02	1.00	0.03	0.01	1.00	0.03	0.08	1.00	0.16	0.04	1.00	0.06
WECS	0.35	0.09	0.14	0.23	0.16	0.19	0.10	0.28	0.14	0.25	0.29	0.26	0.23	0.06	0.09	0.13	0.53	0.21	0.22	0.23	0.16
SD	0.73	0.55	0.63	0.72	0.36	0.48	0.88	0.33	0.49	0.45	0.45	0.48	0.84	0.40	0.54	0.95	0.55	0.70	0.76	0.44	0.55
NCA	0.63	0.66	**0.65**	0.59	0.65	**0.61**	0.71	0.56	**0.63**	0.64	0.51	**0.58**	0.79	0.61	**0.69**	0.85	0.78	**0.82**	0.70	0.63	**0.66**

Table 2. Tag-to-class alignment performance for OSM tags to DBpedia classes.

Name	France			Germany			Great Britain			Spain			USA			Australia			Average		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
CUPID	0.32	1.00	0.48	0.18	1.00	0.31	0.41	1.00	0.58	0.44	1.00	0.63	0.10	1.00	0.17	0.48	1.00	0.65	0.32	1.00	0.47
LD	0.31	0.57	0.41	0.32	0.37	0.34	0.73	0.46	0.57	0.34	0.94	0.50	0.42	0.97	0.59	0.58	0.62	0.60	0.45	0.65	0.50
EMBDI	0.16	1.00	0.28	0.09	1.00	0.17	0.29	1.00	0.45	0.24	1.00	0.38	0.33	1.00	0.51	0.32	1.00	0.50	0.24	1.00	0.38
SF	0.14	1.00	0.27	0.10	1.00	0.18	0.27	1.00	0.42	0.24	1.00	0.39	0.33	1.00	0.50	0.30	1.00	0.46	0.23	1.00	0.37
WECS	0.30	65	0.41	0.16	0.97	0.28	0.22	0.96	0.36	0.38	0.67	0.49	0.41	0.95	0.57	0.45	0.66	0.53	0.32	0.81	0.44
SD	0.92	0.57	0.70	0.34	0.98	0.50	0.57	0.88	0.69	0.83	0.58	0.69	0.70	0.47	0.58	0.95	0.55	0.70	0.71	0.67	0.64
NCA	0.95	0.90	**0.92**	0.96	0.79	**0.87**	0.81	0.84	**0.83**	1.00	0.84	**0.91**	0.70	0.70	**0.70**	0.95	0.76	**0.85**	0.90	0.81	**0.85**

Table 3. Example tag-to-class alignments obtained using the NCA approach.

Wikidata: France	Germany	Great Britain	Russia	USA	Australia
amenity=bicycle_rental: bicycle-sharing station	amenity=cinema: movie theater	railway=station: railway station	station=subway: metro station	landuse=reservoir: reservoir	amenity=library: public library
DBpedia: France	Germany	Great Britain	Spain	USA	Australia
railway=station: Place	place=municipality: Place	place=hamlet: Place	railway=station: ArchitecturalStructure	man_made=lighthouse: Location	public_transport=station: Infrastructure

In contrast, DBpedia indicates three classes per entity (i.e., the specialized and more generic classes at the higher levels of the DBpedia ontology). This property makes the model trained on the DBpedia knowledge graph more confident regarding the generic classes, such that generic classes obtain higher F1-scores than the specialized classes. Our observations indicate that it is desirable to obtain more training examples that align entities with more specific classes, such as in the Wikidata dataset. Table 3 illustrates the most confident tag-to-class alignments in terms of the obtained model activations using the NCA approach. As discussed above, Wikidata alignments with high confidence scores are more specific than those obtained on DBpedia.

5.2 Influence of the Shared Latent Space

Table 4 summarizes the performance of the proposed NCA approach and NCA without the shared latent space for tag-to-class alignment of OSM with Wikidata and DBpedia, respectively. We observe that the shared latent space helps to achieve an increase in F1-score of 34% points and 11% points for Wikidata and DBpedia, respectively. Compared to the Wikidata datasets, we observe smaller improvements on DBpedia datasets. DBpedia has an imbalance between the tags

Table 4. Tag-to-class alignment performance for Wikidata and DBpedia.

Approach	Avg. Wikidata			Avg. DBpedia		
	Precision	Recall	F1	Precision	Recall	F1
NCA w/o shared latent space	0.48	0.25	0.32	0.65	0.88	0.74
NCA	0.70	0.63	**0.66**	0.90	0.81	**0.85**

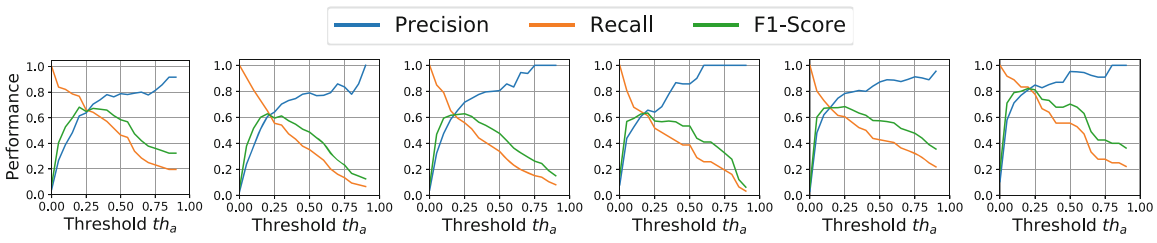
and classes, resulting in many-to-one alignments between tags and classes, where one class corresponds to several tags. For example, in all DBpedia datasets, the *place* and *populatedPlace* are frequently occurring classes for various tags such as `tourism=museum`, `place=village`, `place=town`. In such a case, DBpedia properties add less specific information to the matching process. Furthermore, we observe a high F1-score of the proposed NCA approach without the shared latent space on the DBpedia dataset. Intuitively, further improving these high scores is more difficult than improving the comparably low scores on Wikidata (e.g., 0.32 F1-score on Wikidata). In summary, the shared latent space improves the performance, with the highest improvements on Wikidata.

5.3 Confidence Threshold Tuning

We evaluate the influence of the confidence threshold value th_a on the precision, recall, and F1-score. The threshold th_a indicates the minimum similarity at which we align a tag to a class. Figure 4 and 5 present the alignment performance with respect to th_a for Wikidata and DBpedia. As expected, we observe a general trade-off between precision and recall, whereas higher values of th_a result in higher precision and lower recall. We select the confidence threshold of $th_a = 0.25$ and $th_a = 0.4$ for Wikidata and DBpedia, respectively, as these values allow balancing precision and recall. The threshold can be tuned for specific regions.

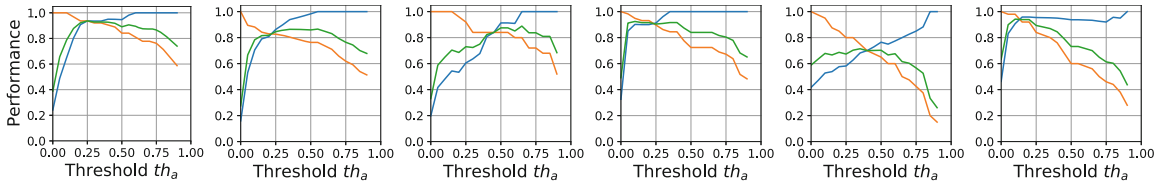
5.4 Alignment Impact

To assess the impact of NCA, we compare the number of OSM entities that can be annotated with semantic classes using the alignment discovery by NCA



(a) France (b) Germany (c) GB (d) Russia (e) USA (f) Australia

Fig. 4. Precision, recall, and F1-score vs. the confidence threshold for Wikidata. (Color figure online)



(a) France (b) Germany (c) GB (d) Spain (e) USA (f) Australia

Fig. 5. Precision, recall, and F1-score vs. the confidence threshold for DBpedia. (Color figure online)

with the number of entities that are linked to a KG in the currently existing datasets. For Wikidata, we observe 2,004,510 linked OSM entities and 10,163,762 entities annotated with semantic classes using NCA. This result corresponds to an increase of 407.04% of entities with semantic class annotations. For DBpedia, we observe 1,396,378 linked OSM entities and 8,301,450 entities annotated with semantic classes using NCA. This result corresponds to an increase of 494.5% of entities with semantic class annotations. We provide the resulting annotations as a part of the WorldKG knowledge graph[8].

6 Related Work

This work is related to ontology alignment, alignment of tabular data, feature space alignment, and link discovery.

Ontology Alignment. Ontology alignment (also ontology matching) aims to establish correspondences between the elements of different ontologies. The efforts to interlink open semantic datasets and benchmark ontology alignment approaches have been driven by the W3C SWEO Linking Open Data community project[9] and the Ontology Alignment Evaluation Initiative (OAEI)[10] [1]. Ontology alignment is conducted at the element-level and structure-level [20]. The element-level alignment typically uses natural language descriptions of the ontology elements, such as labels and definitions. Element-level alignment adopts string similarity metrics such as, e.g., edit distance. Structure-level alignment exploits the similarity of the neighboring ontology elements, including the taxonomy structure, as well as shared instances [17]. Element-level and structure-level alignment have also been adopted to align ontologies with relational data [6] and tabular data [32]. Jiménez-Ruiz et al. [11] divided the alignment task into independent, smaller sub-tasks, aiming to scale up to very large ontologies. In machine learning approaches, such as the GLUE architecture [7], semantic mappings are learned in a semi-automatic way. In [19], a matching system integrates string-based and semantic similarity features. Recently, more complex

[8] WorldKG knowledge graph: http://www.worldkg.org.

[9] https://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData.

[10] OAEI evaluation campaigns: http://oaei.ontologymatching.org.

approaches using deep neural networks have been proposed for ontology alignment and schema matching [3,22,31]. The lack of a well-defined ontology in OSM hinders the application of ontology alignment approaches. In contrast, the instance-based NCA approach enables an effective alignment of tags to classes.

Tabular Data Alignment. Another branch of research investigated the schema alignment of tabular data [23]. EmbDi [5] approach uses random walks and embeddings to find similarities between schema elements. Cupid [13] matches schema elements based on element names, structure, and data types. Similarity Flooding [14] transforms a table into a directed labeled graph in which nodes represent columns to compute similarity values iteratively. We employ the EmbDi, Cupid, and Similarity Flooding algorithms as baselines for our evaluation. Although the conversion of OSM key-value-based data into a tabular form is possible in principle, the resulting tables are highly sparse. Therefore, as seen in Sect. 4.3, tabular data alignment approaches do not perform well on the alignment task addressed in this work.

Feature Space Alignment. Recently, various studies investigated the alignment of feature spaces extracted from different data sources. Application domains include computer vision [8] and machine translation [12]. Ganin et al. [9] proposed a neural domain adaptation algorithm that considers labeled data from a source domain and unlabeled data from a target domain. While this approach was originally used to align similar but different distributions of feature spaces, we adopt the gradient reversal layer proposed in [9] to fuse information from the disjoint features spaces of OSM and KGs, not attempted previously.

Link Discovery. Link Discovery is the task of identifying semantically equivalent resources in different data sources [16]. Nentwig et al. [16] provide a recent survey of link discovery frameworks with prominent examples, including Silk [29] and LIMES [18]. In particular, the Wombat algorithm, integrated within the LIMES framework [24], is a state-of-the-art approach for link discovery in knowledge graphs. Specialized approaches [27] focus on link discovery between OSM and knowledge graphs. We build on existing links between OSM and knowledge graphs to align knowledge graph classes to OSM tags in this work.

7 Conclusion

In this paper, we presented NCA – the first neural approach for tag-to-class alignment between OpenStreetMap and knowledge graphs. We proposed a novel shared latent space that seamlessly fuses features from knowledge graphs and OSM in a joint model and makes them simultaneously accessible for the schema alignment. Our model builds this space as the core part of neural architecture, incorporating an auxiliary classification model and an adversarial component. Furthermore, we proposed an effective algorithm that extracts tag-to-class alignments from the resulting shared latent space with high precision. Our evaluation results demonstrate that NCA is highly effective and outperforms the baselines by up to 37% points F1-score. We make our code and manually annotated ground

truth data publicly available to facilitate further research. We believe that NCA is applicable to other geographic datasets having similar data structure as OSM; we leave such applications to future work.

Acknowledgements. This work was partially funded by DFG, German Research Foundation ("WorldKG", DE 2299/2-1), BMBF, Germany ("Simple-ML", 01IS18054) and BMWi, Germany ("d-E-mand", 01ME19009B).

References

1. Algergawy, A., et al.: Results of the ontology alignment evaluation initiative 2019. In: OM-2019. CEUR Workshop Proceedings, vol. 2536, pp. 46–85 (2019)
2. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: DBpedia: a nucleus for a web of open data. In: Aberer, K., et al. (eds.) ASWC/ISWC -2007. LNCS, vol. 4825, pp. 722–735. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-76298-0_52
3. Bento, A., Zouaq, A., Gagnon, M.: Ontology matching using convolutional neural networks. In: LREC 2020, pp. 5648–5653. ELRA (2020)
4. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Trans. Assoc. Comput. Linguist. **5**, 135–146 (2017)
5. Cappuzzo, R., Papotti, P., Thirumuruganathan, S.: Creating embeddings of heterogeneous relational datasets for data integration tasks. In: SIGMOD 2020, pp. 1335–1349. ACM (2020)
6. Demidova, E., Oelze, I., Nejdl, W.: Aligning freebase with the YAGO ontology. In: CIKM 2013, pp. 579–588. ACM (2013)
7. Doan, A., Madhavan, J., Domingos, P.M., Halevy, A.Y.: Ontology matching: a machine learning approach. In: Staab, S., Studer, R. (eds.) Handbook on Ontologies. International Handbooks on Information Systems, pp. 385–404. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-24750-0_19
8. Fernando, B., Habrard, A., Sebban, M., Tuytelaars, T.: Unsupervised visual domain adaptation using subspace alignment. In: ICCV 2013. IEEE (2013)
9. Ganin, Y., et al.: Domain-adversarial training of neural networks. J. Mach. Learn. Res. **17**, 59:1–59:35 (2016)
10. Gottschalk, S., Demidova, E.: EventKG - the hub of event knowledge on the web - and biographical timeline generation. Semantic Web **10**(6), 1039–1070 (2019)
11. Jiménez-Ruiz, E., Agibetov, A., Chen, J., Samwald, M., Cross, V.: Dividing the ontology alignment task with semantic embeddings and logic-based modules. In: ECAI 2020. FAIA, vol. 325, pp. 784–791. IOS Press (2020)
12. Lample, G., Conneau, A., Ranzato, M., Denoyer, L., Jégou, H.: Word translation without parallel data. In: ICLR 2018. OpenReview.net (2018)
13. Madhavan, J., Bernstein, P.A., Rahm, E.: Generic schema matching with cupid. In: VLDB 2001, pp. 49–58. Morgan Kaufmann (2001)
14. Melnik, S., Garcia-Molina, H., Rahm, E.: Similarity flooding: a versatile graph matching algorithm and its application to schema matching. In: ICDE 2002 (2002)
15. Neis, P.: OSMstats. https://osmstats.neis-one.org/. Accessed 10 Apr 2021
16. Nentwig, M., Hartung, M., Ngomo, A.N., Rahm, E.: A survey of current link discovery frameworks. Semantic Web **8**(3), 419–436 (2017)

17. Ngo, D.H., Bellahsene, Z., Todorov, K.: Opening the black box of ontology matching. In: Cimiano, P., Corcho, O., Presutti, V., Hollink, L., Rudolph, S. (eds.) ESWC 2013. LNCS, vol. 7882, pp. 16–30. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-38288-8_2
18. Ngomo, A.N., Auer, S.: LIMES - a time-efficient approach for large-scale link discovery on the web of data. In: IJCAI 2011, pp. 2312–2317. IJCAI/AAAI (2011)
19. Nkisi-Orji, I., Wiratunga, N., Massie, S., Hui, K., Heaven, R.: Ontology alignment based on word embedding and random forest classification. In: ECML PKDD (2018)
20. Otero-Cerdeira, L., Rodríguez-Martínez, F.J., Gómez-Rodríguez, A.: Ontology matching: a literature review. Expert Syst. Appl. **42**(2), 949–971 (2015)
21. Paulheim, H., Bizer, C.: Type inference on noisy RDF data. In: ISWC 2013 (2013)
22. Qiu, L., Yu, J., Pu, Q., Xiang, C.: Knowledge entity learning and representation for ontology matching based on deep neural networks. Clust. Comput. **20**, 969–977 (2017)
23. Rahm, E., Bernstein, P.A.: A survey of approaches to automatic schema matching. VLDB J. **10**(4), 334–350 (2001)
24. Sherif, M.A., Ngonga Ngomo, A.-C., Lehmann, J.: Wombat – a generalization approach for automatic link discovery. In: Blomqvist, E., Maynard, D., Gangemi, A., Hoekstra, R., Hitzler, P., Hartig, O. (eds.) ESWC 2017. LNCS, vol. 10249, pp. 103–119. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-58068-5_7
25. Stadler, C., Lehmann, J., Höffner, K., Auer, S.: LinkedGeoData: a core for a web of spatial open data. Semantic Web **3**(4), 333–354 (2012)
26. Pellissier Tanon, T., Weikum, G., Suchanek, F.: YAGO 4: a reason-able knowledge base. In: Harth, A., et al. (eds.) ESWC 2020. LNCS, vol. 12123, pp. 583–596. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-49461-2_34
27. Tempelmeier, N., Demidova, E.: Linking OpenStreetMap with knowledge graphs - link discovery for schema-agnostic volunteered geographic information. Future Gener. Comput. Syst. **116**, 349–364 (2021)
28. Unal, O., Afsarmanesh, H.: Using linguistic techniques for schema matching. In: ICSOFT 2006, pp. 115–120. INSTICC Press (2006)
29. Volz, J., Bizer, C., Gaedke, M., Kobilarov, G.: Silk - A link discovery framework for the web of data. In: LDOW 2009. CEUR, vol. 538. CEUR-WS.org (2009)
30. Vrandecic, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Commun. ACM **57**(10), 78–85 (2014)
31. Xiang, C., Jiang, T., Chang, B., Sui, Z.: ERSOM: a structural ontology matching approach using automatically learned entity representation. In: EMNLP (2015)
32. Zhang, S., Balog, K.: Web table extraction, retrieval, and augmentation: a survey. ACM Trans. Intell. Syst. Technol. **11**(2), 13:1–13:35 (2020)

Appendix B

Publication: Iterative Geographic Entity Alignment with Cross-Attention

Iterative Geographic Entity Alignment with Cross-Attention

Alishiba Dsouza[1( )], Ran Yu[1,2], Moritz Windoffer[1], and Elena Demidova[1,2]

[1] Data Science & Intelligent Systems (DSIS), University of Bonn, Bonn, Germany
{dsouza,elena.demidova}@cs.uni-bonn.de, {ran.yu,s5mowind}@uni-bonn.de
[2] Lamarr Institute for Machine Learning and Artificial Intelligence, Bonn, Germany
https://lamarr-institute.org/

Abstract. Aligning schemas and entities of community-created geographic data sources with ontologies and knowledge graphs is a promising research direction for making this data widely accessible and reusable for semantic applications. However, such alignment is challenging due to the substantial differences in entity representations and sparse interlinking across sources, as well as high heterogeneity of schema elements and sparse entity annotations in community-created geographic data. To address these challenges, we propose a novel cross-attention-based iterative alignment approach called IGEA in this paper. IGEA adopts cross-attention to align heterogeneous context representations across geographic data sources and knowledge graphs. Moreover, IGEA employs an iterative approach for schema and entity alignment to overcome annotation and interlinking sparsity. Experiments on real-world datasets from several countries demonstrate that our proposed approach increases entity alignment performance compared to baseline methods by up to 18% points in F1-score. IGEA increases the performance of the entity and tag-to-class alignment by 7 and 8% points in terms of F1-score, respectively, by employing the iterative method.

Keywords: Geographic Knowledge Graph · Iterative Neural Entity Alignment

1 Introduction

Knowledge graphs provide a backbone for emerging semantic applications in the geographic domain, including geographic question answering and point of interest recommendations. However, general-purpose knowledge graphs such as Wikidata [23], DBpedia [14], and YAGO [19] contain only a limited number of popular geographic entities, restricting their usefulness in this context. In contrast, OpenStreetMap (OSM)[1][2] is a community-created world-scale geographic

[1] https://www.openstreetmap.org/.

[2] OpenStreetMap, OSM and the OpenStreetMap magnifying glass logo are trademarks of the OpenStreetMap Foundation, and are used with their permission. We are not endorsed by or affiliated with the OpenStreetMap Foundation.

T. R. Payne et al. (Eds.): ISWC 2023, LNCS 14265, pp. 216–233, 2023.
https://doi.org/10.1007/978-3-031-47240-4_12

data source containing millions of geographic entities. However, the community-driven nature of OSM leads to highly heterogeneous and sparse annotations at both the schema and instance levels, which lack machine-interpretable semantics and limit the accessibility and reusability of OSM data. Knowledge graphs extracted from OSM and dedicated to geographic entities such as LinkedGeoData [1] and WorldKG [7] focus on a selection of well-annotated geographic classes and entities and do not take full advantage of OSM data. Tighter interlinking of geographic data sources with knowledge graphs can open up the rich community-created geographic data sources to various semantic applications.

Interlinking geographic data sources with knowledge graphs is challenging due to the heterogeneity of their schema and entity representations, along with the sparsity of entity annotations and links between sources. Knowledge graphs such as Wikidata adopt ontologies to specify the semantics of entities through classes and properties. Taking the entity Berlin as an example, Table 1a and 1b illustrate its representation in OSM and Wikidata. The property wdt:P31 (instance of) in Wikidata specifies the entity type. In contrast, OSM annotates geographic entities using key-value pairs called tags, often without clear semantics. The distinction of whether a key-value pair represents an entity type or an attribute is not provided. For instance, in Table 1, the key *capital* in OSM corresponds to a binary value specifying whether the location is the capital of a country. In contrast, the Wikidata property wdt:P1376 (*capital of*) is an object property linked to an entity of type country. Moreover, user-defined key-value pairs in OSM lead to highly heterogeneous and sparse annotations, where many entities do not have comprehensive annotations and many key-value pairs are rarely reused. Finally, sparse and often inaccurate interlinking makes training supervised alignment algorithms difficult. As illustrated in the example, the values, such as the geo-coordinates of the same real-world entity Berlin, differ between sources. Such differences in representation, coupled with the heterogeneity and sparsity of OSM annotations and the lack of links, make schema and entity alignment across sources extremely challenging.

Recently, several approaches have been proposed to interlink knowledge graphs to OSM at the entity and schema level, to lift the OSM data into a semantic representation, and to create geographic knowledge graphs [1,6,13,21]. For example, LinkedGeoData [1] relies on manual schema mappings and provides high-precision entity alignment using labels and geographic distance for a limited number of well-annotated classes. OSM2KG [21] – a linking method for geographic entities, embeds the tags of geographic entities for entity representation and interlinking. The NCA tag-to-class alignment [6] enables accurate matching of frequent tags to classes, but does not support the alignment of rare tags. The recently proposed WorldKG knowledge graph [7] incorporates the information extracted by NCA and OSM2KG, but is currently limited to the well-annotated geographic classes and entities. Overall, whereas several approaches for linking geographic entities and schema elements exist, they are limited to well-annotated classes and entities, they rely on a few properties and do not sufficiently address the representation heterogeneity and annotation sparsity.

Table 1. An excerpt of the Berlin representation in OSM and Wikidata.

(a) OSM tags.

Key	Value
name	Berlin
place	city
population	3769962
way	POINT(52.5183 13.4179)
capital	yes

(b) Wikidata triples. `wd:Q64` identifies Berlin.

Subject	Predicate	Object
`wd:Q64`	`rdfs:label` (*label*)	Berlin
`wd:Q64`	`wdt:P31` (*instance of*)	`wd:Q515` (*city*)
`wd:Q64`	`wdt:P1082` (*population*)	3677472
`wd:Q64`	`wdt:P625` (*coordinate location*)	52°31'N, 13°23'E
`wd:Q64`	`wdt:P1376` (*capital of*)	`wd:Q183` (*Germany*)

In this paper, we propose IGEA – a novel iterative geographic entity alignment approach. IGEA relies on a cross-attention mechanism to align heterogeneous context representations across community-created geographic data and knowledge graphs. This model learns the representations of the entities through the tags and properties and reduces the dependency on specific tags and labels. Furthermore, to overcome the annotation and interlinking sparsity problem, IGEA employs an iterative approach for tag-to-class and entity alignment that starts from existing links and enriches the links with alignment results from previous iterations. We evaluate our approach on real-world OSM, Wikidata, and DBpedia datasets. The results demonstrate that, compared to state-of-the-art baselines, the proposed approach can improve the performance of entity alignment by up to 18% points, in terms of F1-score. By employing the iterative method, IGEA increases the performance of the entity and tag-to-class alignment by 7 and 8% points in terms of F1-score, respectively.

In summary, our contributions are as follows:

- We propose IGEA – a novel iterative cross-attention-based approach to interlink geographic entities, bridging the representation differences in community-created geographic data and knowledge graphs.
- To overcome the sparsity of annotations and links, IGEA employs an iterative method for tag-to-class and entity alignment, with integrated candidate blocking mechanisms for efficiency and noise reduction.
- We demonstrate that IGEA substantially outperforms the baselines in F1-score through experiments on several real-world datasets.

2 Problem Statement

In this section, we introduce the relevant concepts and formalize the problem addressed in this paper.

Definition 1 (Knowledge Graph). *A* knowledge graph $KG = (E, C, P, L, F)$ *consists of a set of entities* E*, a set of classes* $C \subset E$*, a set of properties* P*, a set of literals* L *and a set of relations* $F \subseteq E \times P \times (E \cup L)$*.*

Entities of knowledge graph KG with geo-coordinates L_{geo} are referred to as geographic entities E_{geo}.

Definition 2 (Geographic Entity Alignment). *Given an entity n from a geographic data source G ($n \in G$), and a set of geographic entities E_{geo} from a knowledge graph KG, $E_{geo} \subseteq KG$, determine the entity $e \in E_{geo}$ such that $sameAs(n, e)$ holds.*

In the example in Table 1, as a result of the geographic entity alignment, Berlin from OSM will be linked to Berlin from Wikidata with a *sameAs* link.

Definition 3 (Geographic Class Alignment). *Given a geographic data source G and a knowledge graph KG, find a set of pairs of class elements of both sources, such that elements in each pair (s_i, s_j), $s_i \in G$ and $s_j \in KG$, describe the same real-world concept.*

In the example illustrated in Table 1, the tag place=city from OSM will be linked to the *city* (wd:Q515) class of Wikidata.

In this paper, we address the task of geographic entity alignment through iterative learning of class and entity alignment.

3 The *IGEA* Approach

In this section, we introduce the proposed IGEA approach. Figure 1 provides an approach overview. In the first step, IGEA conducts geographic class alignment based on known linked entities between OSM and KG with the NCA approach [6]. The resulting tag-to-class alignment is further adopted for blocking in the candidate generation step. Then IGEA applies the cross-attention-based entity alignment module to the candidate set to obtain new links. IGEA repeats this process iteratively with the resulting high-confidence links for several iterations. In the following, we present the proposed IGEA approach in more detail.

3.1 Geographic Class Alignment

We adopt the NCA alignment approach introduced in [6] to conduct tag-to-class alignment. The NCA approach aligns OSM tags with the KG classes. NCA relies on the linked entities from both sources, OSM and a KG, and trains a neural model to learn the representations of the tags and classes. The NCA model creates the shared latent space while classifying the OSM entities into the knowledge graph classes. NCA then probes the resulting classification model to obtain the tag-to-class alignments. NCA selects all matches above a certain threshold value. After applying NCA, we obtain a set of tag-to-class alignments, i.e., (s_i, s_j), $s_i \in G$, and $s_j \in KG$.

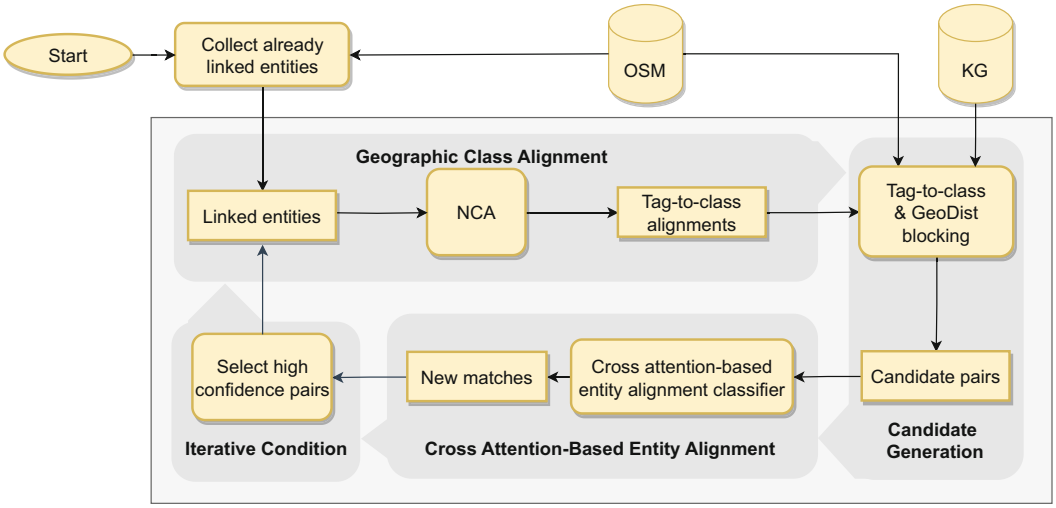


Fig. 1. Overview of the proposed IGEA approach.

3.2 Candidate Generation

OSM contains numerous geographic entities for which we often do not have a match in the KGs. IGEA applies candidate blocking to reduce the search space to make the algorithm more time and complexity efficient. In our task, the objective of the blocking module is to generate a set of candidate entity pairs that potentially match. We built the candidate blocking module based on two strategies, namely entity-type-based and distance-based candidate selection. Entities with a *sameAs* link should belong to the same class. Therefore, we use the tag-to-class alignments produced by the NCA module to select the entities of the same class from both sources to form candidate pairs. Secondly, since we consider only geographic entities, we use spatial distance to reduce the candidate set further and only consider the entities within a threshold distance. Past works observed that a threshold value of around 2000 to 2500 m can work well for most classes [1,13,21]. We choose the threshold of 2500 m as mentioned in [21]. The candidate pairs generated after the candidate blocking step are passed to the cross-attention-based entity alignment module.

3.3 Cross-Attention-Based Entity Alignment

We build a cross-attention-based classification model for entity alignment by classifying a pair of entities into a match or a non-match. Figure 2 illustrates the overall architecture of the entity alignment model. The components of the model are described in detail below.

Entity Representation Module: In this module, we prepare entity representations to serve as the model input. For a given OSM node, we select all tags and create a sentence by concatenating the tags. For a given KG entity, we select all predicates and objects of the entity and concatenate all pairs of predicates and objects to form a sentence. We set the maximum length of a sentence to be input to the model to N_w, where N_w is calculated as the average number of

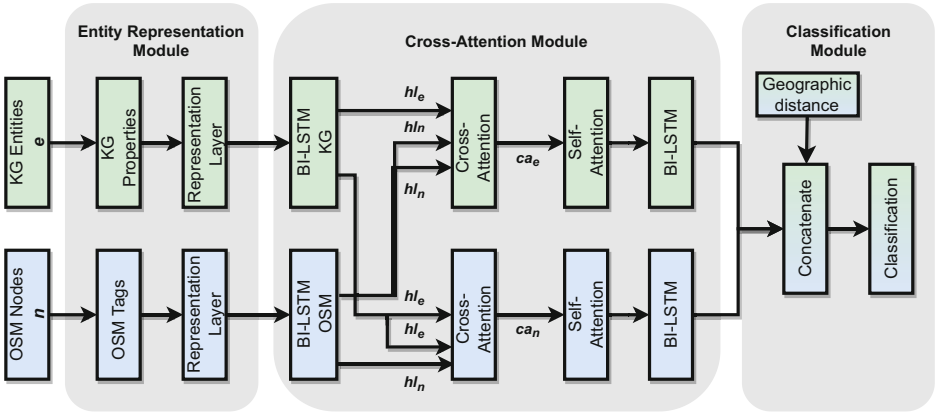


Fig. 2. Cross-attention-based entity alignment model.

words of all entities in the current candidate set. We pass these sentences to the representation layer for each pair of OSM node n and KG entity e.

In the representation layer, the model creates embeddings for the given sentence. We adopt pre-trained fastText word embeddings [3] for the embedding layer. For any word not present in the pre-trained embeddings, we assign a zero vector of size d, where d is the embeddings dimension. In this step, we obtain an array of size $N_w * d$ for each entity.

Cross-Attention Module: We initiate our cross-attention module with a Bi-directional LSTM (BI-LSTM) layer. BI-LSTM models have been demonstrated to perform well on sequential data tasks such as named entity recognition and speech recognition [4,10]. We adopt BI-LSTM since we want the model to learn to answer what comes after a particular key or a property to help the cross-attention layer. We incorporate BI-LSTM layers after the embedding layers for each of the inputs. As an output, the BI-LSTM layer can return the final hidden state or the full sequence of hidden states for all input words. We select the full sequence of hidden states hl_n, hl_e since we are interested in the sequence and not a single output. These sequences of hidden states hl_n, hl_e are then passed to the cross-attention layer.

Cross-Attention Layer: This layer implements the cross-attention mechanism [22] that helps understand the important properties and tags for aligning the entities. As explained in [22], attention scores are built using keys, values, and queries along with their dimensions. For OSM, we adopt the output of the BI-LSTM layer hl_e as key k and query q and hl_n becomes the value v. For KGs, we adopt the output of the BI-LSTM layer hl_n as key k and query q and hl_e becomes the value v. We initialize the weight vectors w_q, w_k, w_v using the Xavier uniform initializer [9]. We then compute the cross-attention weights for OSM as:

$$Q = hl_e * w_q, K = hl_e * w_k, V = hl_n * w_v,$$

$$att = Q \cdot K, att_w = softmax(att), att_c = att_w \cdot V,$$

where att_w is the attention weights and att_c is the context.

Similarly, we compute the attention weights for KGs by interchanging the values of hl_n and hl_e. We then pass the concatenated att_w and att_c as ca_n and ca_e to the self-attention model.

Self-Attention Layer: Adopting both cross-attention and self-attention layers can improve the performance of the models in multi-modal learning [15]. In our case, the intuition behind adopting the self-attention layer is that the model can learn the important tags and properties of a given entity. The formulation of self-attention is similar to that of cross-attention. Instead of using a combination of outputs from the OSM and KG cross-attention layers ca_n and ca_e, we use only one input, either ca_n and ca_e that is the same across k, q, v. We then pass the self-attention output, i.e., concatenated att_w, att_c, through the final layer of Bi-directional LSTM.

Once we have both inputs parsed through all layers, we concatenate the outputs of the Bi-directional LSTM layers along with the distance input that defines the haversine distance between the input entities.

Classification Module: We utilize the linked entities as the supervision for the classification. Each true pair is labeled one, and the remaining pairs generated by the candidate blocking step are labeled zero. The classification layer predicts whether the given pair is a match or not. We pass the concatenated output through a fully connected layer, which is then passed through another fully connected layer with one neuron to predict the final score. We use a sigmoid activation function with binary cross-entropy loss to generate the score for the final match.

3.4 Iterative Geographic Entity Alignment Approach

We create an end-to-end iterative pipeline for aligning KG and OSM entities and schema elements to alleviate the annotation and interlinking sparsity. We apply the IGEA approach at the country level. For a selected country, we collect all entities having geo-coordinates from the KG. In the first iteration, the already linked entities are used as supervision to link unseen entities that are not yet linked. After selecting candidate pairs and classifying them into match and non-match classes, we use a threshold th_a to only select high confidence pairs from the matched class. In the subsequent iterations, we add these high-confidence matched pairs to the linked entities and then run the pipeline starting from NCA-based class alignment again. By doing so, we aim to enhance the performance of entity alignment with tag-to-class alignment-based candidate blocking and tag-to-class alignment with additional newly linked entities. Algorithm 1 provides details of the IGEA approach.

Algorithm 1 The IGEA Algorithm

`Input:`	n, e	OSM and KG linked Entities
	th_a	Alignment threshold
	itr	number of iterations
	con	Country
	kg	KG
`Output:`	$align$	Final entity alignment

$align \Leftarrow \emptyset$
`load`(n, e, con)
$KG_e \Leftarrow$ getCountryEntities(con, kg)
$GT \Leftarrow$getSeedAlignment(con,kg)
while $i < itr$ **do**
  $tag\text{-}to\text{-}class \Leftarrow$ NCA(con,kg,GT)
  $view \Leftarrow$ createView($tag\text{-}to\text{-}class$)
  for all $ent \in KG_e$ **do**
    $candidates \Leftarrow$ `generateCandidates(`ent`,` $view$`, 2500)`
    if $candidates \cap GT \neq \emptyset$ **then**
      $SeenEnt \Leftarrow candidates$
    else
      $UnseenEnt \Leftarrow candidates$
    end if
  end for
  $model \Leftarrow$ classificationModel($seenEnt$)
  $prediction \Leftarrow model(UnseenEnt)$
  for all $pair \in prediction$ **do**
    if $pair_{confidence} > th_a$ **then**
      $align \Leftarrow align \cup \{pair\}$
      $GT \Leftarrow GT \cup \{pair\}$
    end if
  end for
  $i = i + 1$
end while
return $align$

4 Evaluation Setup

This section describes the experimental setup, including datasets, ground truth generation, baselines, and evaluation metrics. All experiments were conducted on an AMD EPYC 7402 24-Core Processor with 1 TB of memory. We implement the framework in Python 3.8. For data storage, we use the PostgreSQL database (version 15.2). We use TensorFlow 2.12.0 and Keras 2.12.0 for neural model building.

4.1 Datasets

For our experiments, we consider OSM, Wikidata, and DBpedia datasets across various countries, including Germany, France, Italy, USA, India, Netherlands,

Table 2. Ground truth size for Wikidata and DBpedia.

	France	Germany	India	Italy	Netherlands	Spain	USA
Wikidata	19082	21165	7001	16584	4427	14145	73115
DBpedia	10921	165	1870	2621	110	4319	14017

and Spain. All datasets were collected in April 2023. For OSM data, we use OSM2pgsql[3] to load the nodes of OSM into the PostgreSQL database. The OSM datasets are collected from GeoFabrik download server[4]. For Wikidata[5] and DBpedia[6], we rely on the SPARQL endpoints. Given a country, we select all entities that are part of the country with property *P17* for Wikidata and *dbo:country* for DBpedia along with geo-coordinates (*P625* for Wikidata and *geo:geometry* for DBpedia).

4.2 Ground Truth

We select the existing links between geographic entities in OSM and KGs as ground truth. Since we consider geographic entities from the already linked entities identified through "wikidata" and "wikipedia" tags, we select entities with geo-coordinates. Table 2 displays the number of ground truth entities for Wikidata and DBpedia knowledge graphs. We consider only those datasets where the number of links in the ground truth data exceeds 1500 to have sufficient data to train the model. For tag-to-class alignment, we use the same ground truth as in the NCA [6] approach.

4.3 Baselines

This section introduces the baselines to which we compare our work, including similarity-based and deep learning-based approaches.

GeoDistance: In this baseline, we select the OSM node for each KG geographic entity so that the distance between the KG entity and the OSM node is the least compared to all other OSM nodes. We consider the distance calculated using the *st_distance* function of PostgreSQL that calculates the minimum geodesic distance as the distance metric.
LGD [1]**:** LinkedGeoData approach utilizes geographic and linguistic distance to match the entities in OSM and KG. Given a pair of geographic entities $e1$ and $e2$, LinkedGeoData considers $\frac{2}{3}ss(e1, e2) + \frac{1}{3}gd(e1, e2) > 0.95$ as a match, where ss is the Jaro-Winkler distance and gd is the logistic geographical distance.

[3] https://osm2pgsql.org/.
[4] https://download.geofabrik.de/.
[5] https://query.wikidata.org/.
[6] https://dbpedia.org/sparql.

Yago2Geo: Yago2Geo [13] considers both string and geographic distance while matching entities by having two filters, one based on Jaro-Winkler similarity (s) between the labels and the second filter based on the Euclidean distance (ed) between the geo-coordinates of the two entities. Given entities $e1$ and $e2$, if $s(e1, e2) > 0.82$ and $ed(e1, e2) < 2000\ meters$, the two entities are matched.

DeepMatcher: DeepMatcher [17] links two entities from different data sources having similar schema. The model learns the similarity between two entities by summarizing and comparing their attribute embeddings. Since our data sources do not follow the same schema, we select the values of keys name, addressCountry, address, and population for OSM. For KGs, we select the values of the equivalent properties label, country, location, and population.

HierMatcher: This baseline [8] aligns entities by jointly matching at token, attribute, and entity levels. At the token level, the model performs the cross-attribute token alignment. At the attribute level, the attention mechanism is applied to select contextually important information for each attribute. Finally, the results from the attribute level are aggregated and passed through fully connected layers that predict the probability of two entities being a match.

OSM2KG: OSM2KG [21] implements a machine learning-based model for the entity alignment between OSM and KGs. The model generated key-value embeddings using the occurrences of the tags and created a feature vector including entity type and popularity of KG entities. We use the default th_{dist} 2500 m and the random forest classification model adopted in the original paper.

OSM2KG-FT: This baseline is a variation of the OSM2KG model where we replace the key-value embeddings of OSM entities with fastText embeddings.

4.4 Evaluation Metrics

The standard evaluation metrics for entity and tag-to-class alignment are precision, recall, and F1-score computed against a reference alignment (i.e., ground truth). We calculate **precision** as the ratio of all correctly identified pairs to all identified pairs. We calculate **recall** as the fraction of all correctly identified pairs to all pairs in the ground truth alignment. **F1-score** is the harmonic mean of recall and precision. The F1-score is most relevant for our analysis since it considers both precision and recall. We use macro averages for the metrics because we have imbalanced datasets in terms of classes.

5 Evaluation

In this section, we discuss the performance of the IGEA model. First, we evaluate the performance of the approach for entity alignment against baselines.

Furthermore, we assess the impact of the number of iterations and thresholds. Finally, we demonstrate the approach effectiveness on unseen entities through a manual assessment. To facilitate the evaluation, we split our data into 70:10:20 for training, validation, and test data with a random seed of 42.

Table 3. Entity alignment performance on the OSM to Wikidata linking.

Name	France			Germany			India			Italy			Netherlands			Spain			USA		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
GeoDist	0.65	0.65	0.65	0.56	0.56	0.56	0.75	0.75	0.75	0.68	0.68	0.68	0.67	0.67	0.67	0.71	0.71	0.71	0.88	0.88	0.88
LGD	0.63	0.61	0.62	0.83	0.81	0.82	0.87	0.68	0.72	0.90	0.68	0.77	0.81	0.79	0.80	0.82	0.40	0.82	0.87	0.84	0.85
Yago2Geo	0.5	0.51	0.50	0.53	0.51	0.50	0.61	0.60	0.60	0.52	0.51	0.50	0.50	0.88	0.64	0.63	0.70	0.65	0.88	0.69	0.73
DeepMatcher	0.62	0.58	0.60	0.74	0.67	0.71	0.77	0.79	0.78	0.89	0.55	0.68	0.83	0.78	0.80	0.87	0.75	0.80	0.93	0.91	0.91
HierarMatch	0.51	0.71	0.59	0.64	0.79	0.70	0.71	0.88	0.79	0.62	0.83	0.71	0.8	0.83	0.81	0.80	0.77	0.78	0.92	0.93	0.92
OSM2KG	0.81	0.79	0.80	0.83	0.82	0.82	0.87	0.81	0.84	0.87	0.79	0.83	0.82	0.69	0.75	0.83	0.82	0.82	0.92	0.81	0.86
OSM2KG-FT	0.83	0.81	0.81	0.89	0.82	0.85	0.91	0.75	0.82	0.89	0.85	0.87	0.89	0.71	0.77	0.88	0.82	0.85	0.95	0.87	0.91
IGEA-1	0.95	0.91	0.94	0.93	0.95	0.94	0.88	0.87	0.87	0.93	0.97	0.94	0.94	0.86	0.90	0.89	0.91	0.90	0.93	0.95	0.94
IGEA-3	0.98	0.99	**0.99**	0.93	0.96	**0.95**	0.96	0.90	**0.93**	0.99	0.97	**0.98**	0.94	0.94	**0.94**	0.98	0.93	**0.95**	0.97	0.97	**0.97**

5.1 Entity Alignment Performance

Tables 3 and 4 present the performance of the IGEA approach and the baselines in terms of precision, recall, and F1-score on the various country datasets for Wikidata and DBpedia knowledge graphs, respectively. IGEA-1 and IGEA-3 indicate the results obtained with the 1st and 3rd iterations of the IGEA approach, respectively. The results demonstrate that the proposed IGEA approach outperforms all the baselines in terms of the F1-score. We achieve up to 18% points F1-score improvement on Wikidata and up to 14% points improvement over DBpedia KGs. IGEA also achieves the best recall and precision on several datasets. Regarding the baselines, as expected, GeoDist performs poorly since the geo-coordinates of the same entity are presented with different precision in OSM and in KGs and are not always in closer proximity to each other. OSM2KG-FT performs the best among the baselines for both KGs. We notice that using the tags with fastText embeddings slightly improves the performance of the OSM2KG over using the occurrence-based key-value embeddings. The deep-learning-based baselines perform on par with the other baselines. The absence of the features such as name and country limits the performance of these deep-learning-based baselines that rely on specific properties. The performance of the name-based baselines such as Yago2Geo and LGD is inconsistent across datasets; a potential reason is the absence of labels in the same language.

Regarding the datasets, the IGEA approach achieved the highest performance improvement on the France and Spain datasets for Wikidata and DBpedia KGs, respectively. The smallest performance improvement over the best-performing baselines is produced on the USA dataset. Data in the USA dataset is mostly in English; furthermore, the USA dataset has the highest percentage of name tags among given countries, which makes string similarity-based baseline

approaches more effective. We notice that India achieves the lowest performance across datasets and KGs. The number of overall properties and tags for entities in India are lower than in other datasets, making IGEA less beneficial. DBpedia results demonstrate better model performance compared to Wikidata. Since DBpedia contains more descriptive properties, it benefits more from employing the cross-attention-based mechanism.

Table 4. Entity alignment performance on the OSM to DBpedia linking.

Name	France			India			Italy			Spain			USA		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
GeoDist	0.39	0.39	0.39	0.35	0.35	0.35	0.58	0.58	0.58	0.40	0.40	0.40	0.64	0.64	0.64
LGD	0.84	0.76	0.79	0.83	0.63	0.72	0.87	0.69	0.76	0.91	0.72	0.78	0.70	0.61	0.64
Yago2Geo	0.70	0.63	0.66	0.67	0.65	0.65	0.73	0.69	0.71	0.73	0.76	0.74	0.54	0.54	0.54
DeepMatcher	0.79	0.85	0.82	0.78	0.85	0.81	0.83	0.73	0.77	0.81	0.73	0.77	0.85	0.86	0.85
HierarMatch	0.69	0.84	0.76	0.73	0.85	0.79	0.66	0.90	0.76	0.55	0.87	0.67	0.81	0.90	0.85
OSM2KG	0.80	0.82	0.80	0.84	0.79	0.81	0.80	0.84	0.81	0.82	0.77	0.79	0.87	0.82	0.84
OSM2KG-FT	0.82	0.87	0.84	0.84	0.82	0.83	0.81	0.89	0.85	0.82	0.82	0.82	0.90	0.91	0.90
IGEA-1	0.92	0.91	0.91	0.89	0.91	0.90	0.95	0.89	0.92	0.96	0.97	0.96	0.97	0.95	0.96
IGEA-3	0.95	0.99	**0.97**	0.96	0.97	**0.97**	0.95	0.98	**0.96**	0.96	0.95	**0.95**	0.99	0.97	**0.98**

Table 5. Ablation study results for the DBpedia datasets.

Name	France			India			Italy			Spain			USA		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
w/o Cross-Attention	0.86	0.81	0.83	0.83	0.82	0.82	0.86	0.77	0.81	0.82	0.81	0.81	0.83	0.84	0.83
w/o Distance	0.85	0.89	0.86	0.81	0.87	0.82	0.81	0.83	0.82	0.79	0.86	0.82	0.82	0.87	0.84
w/o Class-Blocking	0.81	0.93	0.87	0.73	0.94	0.82	0.78	0.93	0.85	0.75	0.92	0.83	0.79	0.96	0.86
IGEA-3	0.95	0.99	**0.97**	0.96	0.97	**0.97**	0.95	0.98	**0.96**	0.96	0.95	**0.95**	0.99	0.97	**0.98**

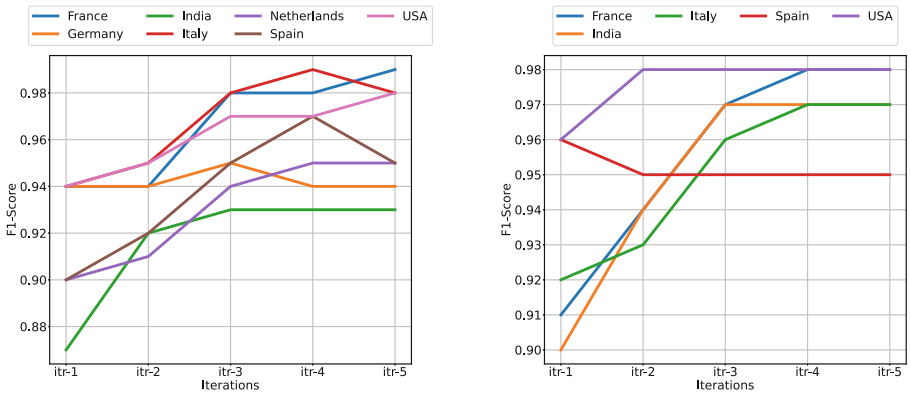
5.2 Ablation Study

Table 5 displays the results of an ablation study to better understand the impact of individual components. We observe that removing the cross-attention layer significantly reduces the performance of the model. The class-based blocking improves the recall but has a sharp decrease in precision, as it creates many noisy matches. Removing geographic distance also results in worse performance compared to the IGEA. The results of the ablation study confirm that the components introduced in the IGEA approach help to achieve the best performance.

5.3 Impact of the Number of Iterations

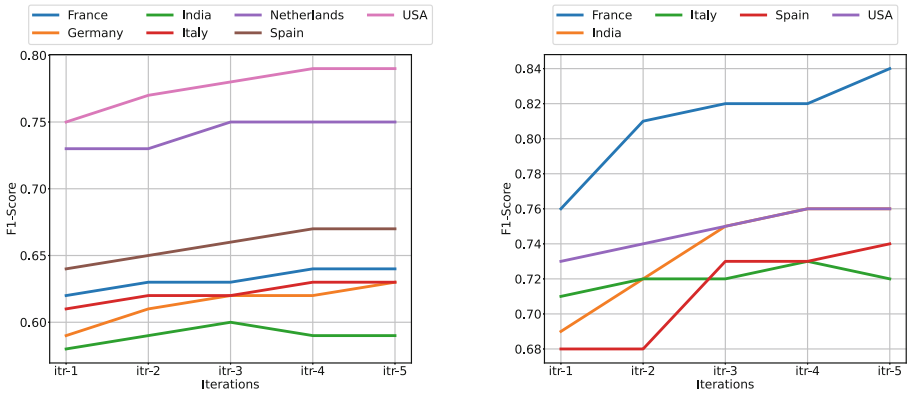
In this section, we evaluate the impact of the number of iterations on the IGEA performance. Figure 3 displays the F1-scores for the entity alignment after each

iteration. We observe that the scores increase in all configurations with the increased number of iterations; after the 3rd iteration, the trend is not continuing. We notice the performance drops for a few countries. After manually checking such drops, we found that the model removes the wrong matches that are part of the ground truth data, which leads to a drop in the evaluation metrics. By adopting an iterative approach, we obtain a maximum improvement of 6 and 7% points in F1-score over Wikidata and DBpedia, respectively. Figure 4 displays the F1-scores for tag-to-class alignment after each iteration. We obtain a maximum increase of 4 and 8% points in the F1-score over Wikidata and DBpedia, respectively. We observe a similar trend as the entity alignment, such that the model performance increases up to the 3rd or 4th iteration. The increased number of aligned tag-class pairs provides more evidence for entity alignment.



(a) Wikidata (b) DBpedia

Fig. 3. Entity alignment performance: F1-scores for 1–5 iterations.



(a) Wikidata (b) DBpedia

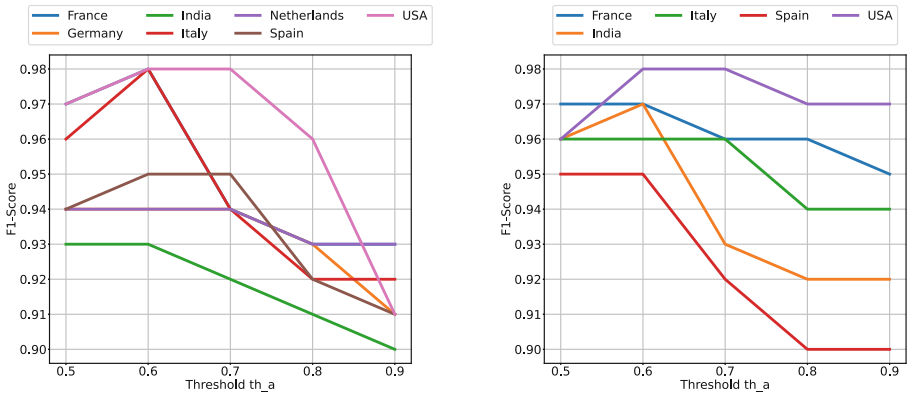
Fig. 4. Tag-to-class alignment performance: F1-scores for 1–5 iterations.

5.4 Alignment Threshold Tuning

We assess the importance of the alignment threshold th_a regarding the F1-score to select the appropriate value of th_a. Figure 5 depicts the F1-scores obtained after the third iteration for threshold values ranging between 0.50 and 0.90 with a gap of 0.1. Overall, the model performs well for all threshold values. Comparing the performance of different th_a values, the highest F1-score is achieved with a $th_a = 0.60$ for both KGs across all datasets. Therefore, in the experiments in other parts of this paper, we set th_a to 0.6.

5.5 Manual Assessment of New Links

We manually assess the quality of the links obtained on unseen data. We create the unseen dataset by considering the entities of Wikidata that are tagged with the country Germany and have a geo-coordinate, but are not present in the ground truth links. We randomly select 100 entities from all iterations and manually verify the correctness of the links. Out of 100 matches, we obtained 89 correct matches. We observe that 6 of the wrong matches are mostly located closer to each other or contained in one another. These entities contain similar property and tag values, making it difficult for the model to understand the difference. For example, Wikidata entity *Q1774543 (Klingermühle)* is contained in OSM node *114219911 (Bessenbach)*. The lack of an English label also hinders the performance. Meanwhile, we observed that IGEA discovers new links between entities and corrects the previously wrong-linked entities. OSM node *1579461216 (Beuel-Ost)* has a Wikidata tag as *Q850834 (Beuel-Mitte)* but using IGEA, the correct Wikidata entity *Q850829 (Beuel-Ost)* has been linked to the OSM node. The performance of the unseen entities demonstrates the effectiveness of the proposed IGEA approach.



(a) Wikidata (b) DBpedia

Fig. 5. Entity alignment performance in terms of F1-Score with different threshold values.

6 Related Work

This section discusses related work in geographic entity alignment, ontology alignment, and iterative learning.

Geographic entity alignment aims to align geographic entities across different geographic sources that refer to the same real-world object. In the past, approaches often relied on geographic distance and linguistic similarity between the labels of the entities [1,13]. LIMES [20] relies on rules to rate the similarity between entities and uses these rules in a supervised model to predict the links. Tempelmeier et al. [21] proposed the OSM2KG algorithm – a machine-learning model to learn a latent representation of OSM nodes and align them with knowledge graphs. OSM2KG also uses KG features such as name, popularity, and entity type to produce more precise links. Recently, deep learning-based models have gained popularity for the task of entity alignment on tabular data. DeepMatcher [8] and HierMatcher [17] use an embedding-based deep learning approach for predicting the matches for tabular datasets. Peeters et al. [18] use contrastive learning with supervision to match entities in small tabular product datasets. In contrast, IGEA adopts the entire entity description, including KG properties and OSM tags, to enhance the linking performance.

Ontology and schema alignment refer to aligning elements such as classes, properties, and relations between ontologies and schemas. Such alignment can be performed at the element and structural levels. Many approaches have been proposed for tabular and relational data schema alignment and rely on the structural and linguistic similarity between elements [5,12,16,26]. Lately, deep learning methods have also gained popularity for the task of schema alignment [2]. Due to the OSM schema heterogeneity and flatness, applying these methods to OSM data is difficult. Recently, Dsouza et al. [6] proposed the NCA model for OSM schema alignment with knowledge graphs using adversarial learning. We adopt NCA as part of the proposed IGEA approach.

Iterative learning utilizes the results of previous iterations in the following iterations to improve the performance of the overall task. In knowledge graphs, iterative learning is mainly adopted in reasoning and completion tasks. Many approaches exploit rule-based knowledge to generate knowledge graph embeddings iteratively. These embeddings are then used for tasks such as link prediction [11,27]. Zhu et al. [28] developed a method for entity alignment across knowledge graphs by iteratively learning the joint low-dimensional semantic space to encode entities and relations. Wang et al. [24] proposed an embedding model for continual entity alignment in knowledge graphs based on latent entity representations and neighbors. In cross-lingual entity alignment, Xie et al. [25] created a graph attention-based model. The model iteratively and dynamically updates the attention score to obtain cross-KG knowledge. Unlike knowledge graphs, OSM does not have connectivity between entities. Therefore, the aforementioned methods are not applicable to OSM. In IGEA, we employ class and entity alignment iteratively to alleviate the data heterogeneity as well as annotation and interlinking sparsity to improve the results of the geographic entity and schema alignment.

7 Conclusion

In this paper, we presented IGEA – a novel iterative approach for geographic entity alignment based on cross-attention. IGEA overcomes the differences in entity representations between community-created geographic data sources and knowledge graphs by using a cross-attention-based model to align heterogeneous context information and predict identity links between geographic entities. By iterating schema and entity alignment, the IGEA approach alleviates the annotation and interlinking sparsity of geographic entities. Our evaluation results on real-world datasets demonstrate that IGEA is highly effective and outperforms the baselines by up to 18% points F1-score in terms of entity alignment. Moreover, we observe improvement in the results of tag-to-class alignment. We make our code publicly available to facilitate further research[7].

Supplemental Material Statement: Sect. 4 provides details for baselines and datasets. Source code, instructions on data collection, and for repeating all experiments are available from GitHub (see footnote 7).

Acknowledgements. This work was partially funded by the DFG, German Research Foundation ("WorldKG", 424985896), the Federal Ministry for Economic Affairs and Climate Action (BMWK), Germany ("ATTENTION!", 01MJ22012C), and DAAD/BMBF, Germany ("KOALA", 57600865).

References

1. Auer, S., Lehmann, J., Hellmann, S.: LinkedGeoData: adding a spatial dimension to the web of data. In: Bernstein, A., et al. (eds.) ISWC 2009. LNCS, vol. 5823, pp. 731–746. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-04930-9_46
2. Bento, A., Zouaq, A., Gagnon, M.: Ontology matching using convolutional neural networks. In: Proceedings of the 12th Language Resources and Evaluation Conference, pp. 5648–5653. ELRA (2020)
3. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Trans. Assoc. Comput. Linguistics **5**, 135–146 (2017)
4. Chiu, J.P.C., Nichols, E.: Named entity recognition with bidirectional LSTM-CNNs. Trans. Assoc. Comput. Linguistics **4**, 357–370 (2016)
5. Demidova, E., Oelze, I., Nejdl, W.: Aligning freebase with the YAGO ontology. In: 22nd ACM International Conference on Information and Knowledge Management, pp. 579–588 (2013)
6. Dsouza, A., Tempelmeier, N., Demidova, E.: Towards neural schema alignment for OpenStreetMap and knowledge graphs. In: Hotho, A., et al. (eds.) ISWC 2021. LNCS, vol. 12922, pp. 56–73. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-88361-4_4
7. Dsouza, A., Tempelmeier, N., Yu, R., Gottschalk, S., Demidova, E.: Worldkg: a world-scale geographic knowledge graph. In: CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, pp. 4475–4484. ACM (2021)

[7] https://github.com/alishiba14/IGEA.

8. Fu, C., Han, X., He, J., Sun, L.: Hierarchical matching network for heterogeneous entity resolution. In: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence,. pp. 3665–3671 (2020)
9. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, pp. 249–256 (2010)
10. Graves, A., Jaitly, N., Mohamed, A.: Hybrid speech recognition with deep bidirectional LSTM. In: Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 273–278 (2013)
11. Guo, S., Wang, Q., Wang, L., Wang, B., Guo, L.: Knowledge graph embedding with iterative guidance from soft rules. In: Proceedings of the 22nd AAAI Conference on Artificial Intelligence, pp. 4816–4823 (2018)
12. Jiménez-Ruiz, E., Agibetov, A., Chen, J., Samwald, M., Cross, V.: Dividing the ontology alignment task with semantic embeddings and logic-based modules. In: Proceedings of the 24th European Conference on Artificial Intelligence, pp. 784–791. FAIA, IOS Press (2020)
13. Karalis, N., Mandilaras, G., Koubarakis, M.: Extending the YAGO2 knowledge graph with precise geospatial knowledge. In: Ghidini, C., et al. (eds.) ISWC 2019. LNCS, vol. 11779, pp. 181–197. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-30796-7_12
14. Lehmann, J., et al.: Dbpedia - a large-scale, multilingual knowledge base extracted from wikipedia. Semantic Web **6**(2), 167–195 (2015)
15. Li, P., et al.: Selfdoc: self-supervised document representation learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5652–5660 (2021)
16. Madhavan, J., Bernstein, P.A., Rahm, E.: Generic schema matching with cupid. In: Proceedings of the 27th International Conference on Very Large Data Bases, pp. 49–58. Morgan Kaufmann (2001)
17. Mudgal, S., et al.: Deep learning for entity matching: a design space exploration. In: Proceedings of the 2018 International Conference on Management of Data, pp. 19–34. ACM (2018)
18. Peeters, R., Bizer, C.: Supervised contrastive learning for product matching. In: Companion of the Web Conference 2022, pp. 248–251. ACM (2022)
19. Rebele, T., et al.: YAGO: a multilingual knowledge base from wikipedia, wordnet, and geonames. In: Proceedings of the 15th International Semantic Web Conference, pp. 177–185 (2016)
20. Sherif, M.A., Ngomo, A.N., Lehmann, J.: Wombat - a generalization approach for automatic link discovery. In: Proceedings of the 14th Extended Semantic Web Conference, pp. 103–119 (2017)
21. Tempelmeier, N., Demidova, E.: Linking OpenStreetMap with knowledge graphs - link discovery for schema-agnostic volunteered geographic information. Future Gener. Comput. Syst. **116**, 349–364 (2021)
22. Vaswani, A., et al.: Attention is all you need. In: Proceedings of the Annual Conference on Neural Information Processing Systems, pp. 5998–6008 (2017)
23. Vrandecic, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Commun. ACM **57**(10), 78–85 (2014)
24. Wang, Y., et al.: Facing changes: continual entity alignment for growing knowledge graphs. In: Sattler, U., et al. The Semantic Web - ISWC 2022. ISWC 2022. LNCS, vol. 13489, pp. 196–213. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-19433-7_12

25. Xie, Z., Zhu, R., Zhao, K., Liu, J., Zhou, G., Huang, J.X.: Dual gated graph attention networks with dynamic iterative training for cross-lingual entity alignment. ACM Trans. Inf. Syst. **40**(3), 44:1–44:30 (2022)
26. Zhang, S., Balog, K.: Web table extraction, retrieval, and augmentation: a survey. ACM Trans. Intell. Syst. Technol. **11**(2), 13:1–13:35 (2020)
27. Zhang, W., et al.: Iteratively learning embeddings and rules for knowledge graph reasoning. In: Proceedings of the World Wide Web Conference, pp. 2366–2377. ACM (2019)
28. Zhu, H., Xie, R., Liu, Z., Sun, M.: Iterative entity alignment via joint knowledge embeddings. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence, pp. 4258–4264 (2017)

Appendix C

Publication: WorldKG: A World-Scale Geographic Knowledge Graph

This is the author's version of the accepted paper. The original version of record can be found at: https://dl.acm.org/doi/abs/10.1145/3459637.3482023

WorldKG: A World-Scale Geographic Knowledge Graph

Alishiba Dsouza
dsouza@cs.uni-bonn.de
Data Science & Intelligent Systems
University of Bonn

Nicolas Tempelmeier
tempelmeier@L3S.de
L3S Research Center
Leibniz Universität Hannover

Ran Yu
ran.yu@uni-bonn.de
Data Science & Intelligent Systems
University of Bonn

Simon Gottschalk
gottschalk@L3S.de
L3S Research Center
Leibniz Universität Hannover

Elena Demidova
elena.demidova@cs.uni-bonn.de
Data Science & Intelligent Systems
University of Bonn

ABSTRACT

OpenStreetMap is a rich source of openly available geographic information. However, the representation of geographic entities, e.g., buildings, mountains, and cities, within OpenStreetMap is highly heterogeneous, diverse, and incomplete. As a result, this rich data source is hardly usable for real-world applications. This paper presents WorldKG - a new geographic knowledge graph aiming to provide a comprehensive semantic representation of geographic entities in OpenStreetMap. We describe the WorldKG knowledge graph, including its ontology that builds the semantic dataset backbone, the extraction procedure of the ontology and geographic entities from OpenStreetMap, and the methods to enhance entity annotation. We perform statistical and qualitative dataset assessment, demonstrating the large scale and high precision of the semantic geographic information in WorldKG.

CCS CONCEPTS

• **Information systems** → *Information integration.*

KEYWORDS

Knowledge Graph; OpenStreetMap; Semantic Geospatial Data

Resource type: Dataset
Website and documentation: http://www.worldkg.org
Dataset DOI: https://zenodo.org/record/4953986

1 INTRODUCTION

OpenStreetMap (OSM) is a rich source of openly available volunteered geographic information, including over 6.8 billion geographic entities in 188 countries contributed by over 7.6 million volunteers [23]. OSM is adopted in a variety of real-world applications on the Web and beyond, including map tile generation [15] and routing [17]. However, representations of geographic entities in OSM are highly diverse, including few mandatory properties and numerous heterogeneous tags, i.e., user-defined key-value pairs. The tag-based structure of OSM data does not follow a well-defined ontology, significantly limiting automatic interpretation and use of OSM data in real-world applications.

Knowledge graphs (KGs) have recently emerged as a key technology to provide semantic machine-interpretable information on real-world entities at scale. However, popular general-purpose knowledge graphs such as DBpedia and Wikidata lack coverage of geographic entities [32]. In contrast, specialized geographic knowledge graphs such as LinkedGeoData [3] and YAGO2geo [19] lack coverage of geographic classes. To provide a comprehensive source of semantic geographic information at scale, semantic information in knowledge graphs and community-created geographic sources such as OSM should be tightly integrated and fused.

Integration of OSM and knowledge graphs is inherently difficult. Although some community-defined links between OSM entities and knowledge graphs like Wikidata exist at the instance level, these links are still sparse and cover only certain entity types. As of April 2021, only 0.52% of OSM entities provided links to Wikidata. In our previous work, we proposed initial approaches for the integration of OSM and knowledge graphs at the schema [12], and instance levels [32]. In this work, we build upon the Neural Class Alignment (NCA) approach [12] to provide semantic annotations to OSM entities. Overall, further research efforts are required to facilitate tighter integration and fusion of OSM and knowledge graphs.

This paper presents WorldKG – a novel comprehensive geographic knowledge graph built from the OSM dataset. We create a novel WorldKG ontology by converting the flat OSM schema into a hierarchical ontology structure. The current release of WorldKG V1.0 in June 2021 contains over 100 million geographic entities from 188 countries and over 800 million triples. Overall, the number of geographic entities in WorldKG is two orders of magnitude higher than in Wikidata and DBpedia knowledge graphs. To facilitate the adoption of WorldKG in semantic applications, we align the WorldKG ontology with the Wikidata and DBpedia ontologies using the NCA approach proposed in our previous work [12]. Our evaluation results demonstrate that the alignment enables us to determine correct Wikidata and DBpedia ontology classes of WorldKG entities with over 99% accuracy, on average.

The scale and accuracy of WorldKG can facilitate the broader adoption of semantic geographic knowledge in a variety of real-world applications. Examples include event-centric [9] and geospatial [28] question answering, geographic information retrieval [29], and other cross-domain semantic data-driven applications.

Overall, our main contributions in this paper are as follows:

- We present WorldKG – a new knowledge graph containing large-scale semantic geographic data extracted from OSM.
- We present the WorldKG ontology, which semantically describes geographic entities and links them to the specific classes in the Wikidata and DBpedia ontologies.
- We provide access to WorldKG through a SPARQL endpoint and provide downloadable data files in the standard RDF turtle format [6].
- To ensure reproducibility, we make the source code of the whole pipeline for WorldKG creation publicly available on GitHub under an open MIT license.

The rest of the paper is organized as follows: In Section 2, we discuss the relevance and the expected impact of the proposed WorldKG knowledge graph. Then, we provide formal definitions of an OSM corpus and a knowledge graph in Section 3. We introduce the proposed WorldKG ontology in Section 4 and explain the WorldKG creation process in Section 5. We present the statistics and evaluation results of WorldKG in Section 6. In Section 7, we describe the availability, utility, and sustainability aspects of our dataset. Section 8 provides a real-world application example using WorldKG. We discuss related work in Section 9. Finally, in Section 10, we provide concluding remarks.

2 RELEVANCE AND EXPECTED IMPACT

This section discusses the expected impact of the proposed WorldKG knowledge graph and its significance to the community, applications, and technology adoption.

Relevance to the information and knowledge management community. Large-scale volunteered geographic information has facilitated many widely used applications such as routing services and data visualizations[1]. Nevertheless, due to the data heterogeneity, the potential of such collectively created knowledge is not yet fully exploited. By integrating heterogeneous OSM data using semantic technologies, we construct and maintain a large-scale knowledge graph that consistently represents geographic data originating from different sources and links this data to the relevant entity types in cross-domain knowledge graphs. WorldKG constitutes a geographic data source of semantic representations with high connectivity, interoperability, and accessibility. In the Semantic Web community context, WorldKG provides richer information of geographic entities than the existing cross-domain knowledge graphs. Thus, WorldKG can support the development of various applications, including geographic question answering and information retrieval, point of interest recommendation, and other cross-domain semantic data-driven applications.

Relevance for OpenStreetMap applications. Currently, routing and navigation services such as Useful Maps 2[2] and Baidu Maps[3], and visualization tools based on geographic information (e.g., weather map[4]) are utilizing OSM. Meanwhile, geographic entities in cross-domain knowledge graphs have been used for entity relation referencing, question answering, and other tasks. However, on the one hand, OSM lacks contextual information on its nodes; on the other hand, cross-domain knowledge graphs are not well-populated with up-to-date geographic information. For these reasons, the gaps between OSM and knowledge graphs persist, and the potential of applications utilizing either type of information is substantially limited. By linking OSM nodes to the classes and entities in cross-domain knowledge graphs, WorldKG provides rich contextual information of the geographic entities, which can be used to enhance the existing services. For instance, enriching maps can provide more detailed location information and interconnect different information types (e.g., locations, weather, and events).

Impact on the adoption of Semantic Web technologies. By following best practices in data publishing and maintenance, we ensure the availability and the extensibility of WorldKG. By adopting Semantic Web technologies and standards, the accessibility and reusability of OpenStreetMap data are largely improved, and the effort associated with reusing this data is reduced significantly. With a commitment to maintaining regular updates, we ensure the sustainability of WorldKG. We believe that WorldKG can benefit researchers in various research fields. Examples include geographic and semantic data management, geographic information retrieval, and recommendation. Furthermore, WorldKG can accelerate the development and enhancement of various services, including interactive maps, smart assistants, and geographic recommender systems.

3 OSM AND KNOWLEDGE GRAPHS

WorldKG targets the integration of OpenStreetMap and knowledge graphs. In this section, we briefly describe both data structures and their interlinking. In the context of this work, we refer to the entities with geographic extent, i.e., the entities located on the globe, as geographic entities.

3.1 OpenStreetMap

OpenStreetMap is one of the essential sources of openly available volunteered geographic information globally, including contributions from over 7.6 million volunteers (as of June 2021). OSM captures a vast and continuously growing number of geographic entities, currently counting more than 6.8 billion in 188 countries [23]. The essential components of the OSM data model are nodes, ways, and relations. Nodes represent entities with a geographic point location (e.g., mountain peaks and trees). Ways represent geographic entities of a linear form (e.g., rivers and roads). Relations are groups of elements consisting of nodes, ways, and other relations (e.g., boundaries and bus routes). For the current scope of the WorldKG knowledge graph (WorldKG V1.0), we only consider OSM nodes.

OSM does not follow a strict schema but provides a set of guidelines[5] for volunteers to create and annotate geographic entities. As a result, OSM has a rich and diverse schema with over 80 thousand distinct keys and numerous values.

We formally define the concept of an OSM corpus as follows:

Definition 3.1. An *OSM corpus* $C = (N, T)$ consists of a set of nodes N representing geographic entities, and a set of tags T. Each tag $t \in T$ is represented as a key-value pair, with the key $k \in K$ and a value $v \in V$: $t = \langle k, v \rangle$. A node $n \in N$, $n = \langle i, l, T_n \rangle$ is represented

[1] OSM-based services: https://wiki.openstreetmap.org/wiki/List_of_OSM-based_services
[2] Useful Maps 2: https://map.atownsend.org.uk/maps/map/map.html
[3] Baidu Maps: http://j.map.baidu.com/1CWxF
[4] Weather map: https://maps.darksky.net/
[5] OSM "How to map a": https://wiki.openstreetmap.org/wiki/How_to_map_a

as a tuple containing an identifier i, a geographic location l, and a set of tags $T_n \subset T$.

OSM nodes have a unique identifier and contain various key-value pairs called tags. The following example of the "Zugspitze", the highest mountain of Germany, illustrates the tag structure.

Key	Value
id	27384190
`name`	*Zugspitze*
`natural`	*peak*
`summit:cross`	*yes*
`ele`	2962

Here, the tags with keys such as *summit:cross*, *name* and *ele* (elevation above sea level) serve as properties of the entity, whereas the tag `natural=peak` represent the entity type (in this case equivalent to the DBpedia class `dbo:Mountain`).

3.2 Knowledge Graphs

Knowledge graphs are a rich source of semantic information, containing entities, classes, properties, literals, and relations.

Definition 3.2. A *knowledge graph* $\mathcal{KG} = (E, C, P, L, F)$ consists of a set of entities E, a set of classes $C \subset E$, a set of properties P, a set of literals L, and a set of relations $F \subseteq E \times P \times (E \cup L)$.

Entities in E represent real-world entities and semantic classes. In the context of this work, we are particularly interested in geographic entities in a knowledge graph. Properties in P represent relations connecting two entities, or an entity and a literal value. An entity in a KG can belong to one or more classes, and is typically linked to a class using `rdf:type` or an equivalent property.

Definition 3.3. The class of the entity $e \in E$ in the knowledge graph $\mathcal{KG} = (E, C, P, L, F)$ is denoted as: $class(e) = \{c \in C \mid (e, \texttt{rdf:type}, c) \in F\}$.

The data in a knowledge graph is typically represented in the RDF[6] format having a *subject - predicate - object* structure. Consider the corresponding excerpt from the representation of the entity "Zugspitze" in Wikidata:

Subject	Predicate	Object
*Q*3375	*label*	*Zugspitze*
*Q*3375	*instance of*	*mountain*
*Q*3375	*coordinate*	$47°25'N, 10°59'E$
*Q*3375	*parentpeak*	*Q*15127

In this example, the statement "Q3375 instance of mountain" indicates that the entity belongs to the Wikidata class "mountain".

3.3 Linking OpenStreetMap and KGs

Although OSM contains a vast amount of geospatial data, OSM keys and tags are heterogeneous, do not possess any machine-readable semantics, and are not directly accessible for semantic applications. Knowledge graphs such as Wikidata, DBpedia, and YAGO provide rich ontologies but lack geographic coverage. For instance, in June 2021, 931,574 nodes with the tag `amenity=restaurant` were present in OSM, whereas Wikidata included only 4,391 entities for the equivalent class "restaurant".

Equivalence links between OSM tags and knowledge graph classes are rarely present. Out of around 80,000 OSM keys, only 0.7% are mapped to Wikidata classes. At the ontology level, the alignment is limited by the structural mismatch between the flat OSM schema and hierarchical KG ontologies. Due to the reasons above, the fusion of OSM and KG entities to create a comprehensive semantic geospatial resource is a challenging task.

Table 1: List of prefixes and namespaces used by WorldKG.

Prefix	Namespace
dcterms	http://purl.org/dc/terms/
geo	http://www.opengis.net/ont/geosparql#
osmn	https://www.openstreetmap.org/node/
owl	http://www.w3.org/2002/07/owl#
rdf	http://www.w3.org/1999/02/22-rdf-syntax-ns#
rdfs	http://www.w3.org/2000/01/rdf-schema#
sf	http://www.opengis.net/ont/sf#
uom	http://www.opengis.net/def/uom/OGC/1.0/
wd	http://www.wikidata.org/wiki/
wkg	http://www.worldkg.org/resource/
wkgs	http://www.worldkg.org/schema/

4 WORLDKG ONTOLOGY

The purpose of WorldKG is to provide a comprehensive geospatial knowledge graph by integrating various data sources. We consider the following goals while building the WorldKG ontology:

- To capture geospatial entities in WorldKG.
- To include relations between classes of existing knowledge graphs and WorldKG classes.
- To lift the OSM schema into a hierarchical ontology.
- To provide provenance information for all WorldKG entities.
- To allow for easy extensions of the WorldKG ontology.

We define the WorldKG ontology based on key-value pairs of the OSM schema. Figure 1 presents the WorldKG ontology. Each class in the WorldKG ontology is a subclass of `wkgs:WKGObject`, where the namespace `wkgs` represents WorldKG schema elements (for a list of prefixes and namespaces in WorldKG, see Table 1). WorldKG properties are modeled as `wkgs:WKGProperty` and provide information on OSM tags that do not indicate a type assignment.

Geospatial support. To enable geographic queries on the dataset, we utilize the GeoSPARQL framework proposed by the Open Geospatial Consortium[7]. To provide information about its geographic extent, each `wkgs:WKGObject` entity can be related to a `geo:SpatialObject` via the property `wkgs:spatialObject`, where a `geo:SpatialObject` can be a point, line string or polygon. `geo:SpatialObject` enables the computation of geospatial functions in SPARQL queries (e.g., distance, nearest neighbors). For an example of a query using these functions, see Section 8.

[6] RDF: https://www.w3.org/RDF/

[7] GeoSPARQL: https://www.ogc.org/standards/geosparql

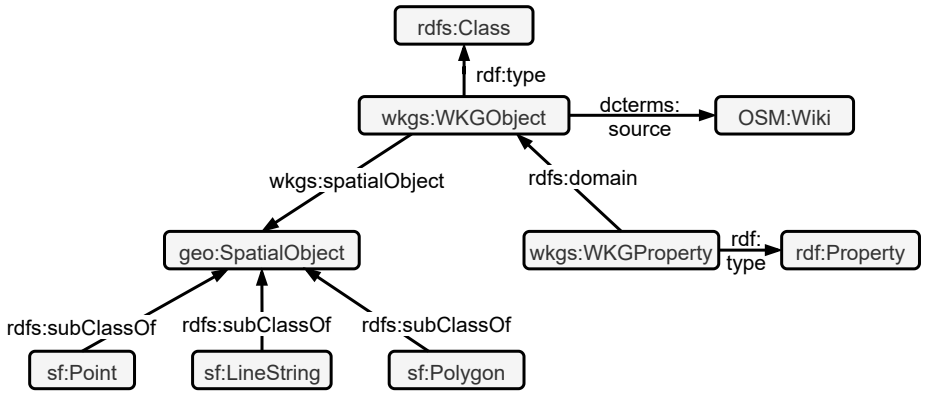


Figure 1: The WorldKG Ontology.

Table 2: Example mappings between OSM tags and Wikidata classes.

OSM tag	Wikidata class (English label)
`natural=peak`	Q8502 (mountain)
`natural=saddle`	Q133056 (mountain pass)
`railway=halt`	Q55678 (railway stop)
`railway=station`	Q55488 (railway station)
`railway=tram_stop`	Q22808404 (station located on surface)
`building=church`	Q16970 (church building)

4.1 WorldKG Classes and Properties

The OSM community provides a list of established key-value pairs as the so-called map feature list[8]. An example of a map feature is the key-value pair `natural=cave_entrance` used to annotate cave entrances in OSM. We use the map feature list to construct a class hierarchy. In particular, we consider all keys in the feature map list as top-level classes (e.g., `natural`). Values assigned to the keys are represented as their subclasses. For example, `cave_entrance` is a subclass of `natural`.

Figure 2 illustrates how the key-value pair `natural=cave_entrance` is represented in the WorldKG ontology.

- The OSM key `natural` is converted into the top-level class `wkgs:Natural`, which summarizes nature entities.
- The OSM value `cave_entrance` is a subclass of `wkgs:Natural`, namely `wkgs:CaveEntrance` representing cave entrances.

We only consider categorical values as subclasses in WorldKG. Other value types, e.g., boolean or numerical values, are not considered as a subclass. Instead, we use the top-level class provided by the corresponding key. For example, an entity with a tag *building=yes* is typed as `wkgs:Building`.

We create the properties from OSM keys that have a valid English OSM Wiki page[9] and are not mapped to own classes. In the example given in Figure 2, `wkgs:addrCountry` is inferred from a key that provides the country in which an entity is located. Each class and property is linked to an OSM Wiki page via `dcterms:source`.

4.2 Schema Alignment with Existing KGs

To link the WorldKG ontology to other existing ontologies, we determine equivalent OSM tags and classes of the Wikidata and DBpedia knowledge graphs. We utilize the Neural Class Alignment (NCA) approach proposed in our previous work [12] to obtain the alignments between OSM tags and the classes of established knowledge graphs. NCA is a 2-step unsupervised machine learning approach. In the first step, we train a supervised neural classification model that learns to classify OSM entities into the respective knowledge graph classes based on their tags (i.e., keys and values). After the training process is completed, we probe the resulting classification model with one tag at a time and get the class activation from the model output layer. Finally, we link the class and tag combinations for which the class activation exceeds an acceptance threshold th_a. The detailed description of the NCA approach is provided in [12].

We train individual models for Wikidata and DBpedia knowledge graphs. We set the acceptance threshold of NCA at $th_a = 0.25$ and $th_a = 0.4$ for Wikidata and DBpedia, respectively. To ensure the quality of tag-to-class matches in WorldKG, we manually verify the resulting matches and discard any wrongly mapped pairs. Table 2 shows example mappings between OSM tags and Wikidata classes obtained using this approach. The alignments between the WorldKG classes and the Wikidata and DBpedia classes are represented using the `owl:equivalentClass` property as shown in the Figure 2.

4.3 Geographic Entity Example

Listing 1 illustrates an example entity description file in *.ttl* format. It contains type information (`wkgs:Restaurant`) and various properties, including its label and opening hours. Via the property `wkgs:spatialObject`, the entity is linked to its respective `geo:SpatialObject`. The `geo:SpatialObject` represents the type of geometry of the entity (`sf:Point`) and the coordinates of the geometry. For each entity, we also provide the property `wkgs:osmLink` that links the entity to the original OSM node.

[8] OSM map feature list: https://wiki.openstreetmap.org/wiki/Map_features
[9] OSM Wiki: https://wiki.openstreetmap.org/wiki/Main_Page

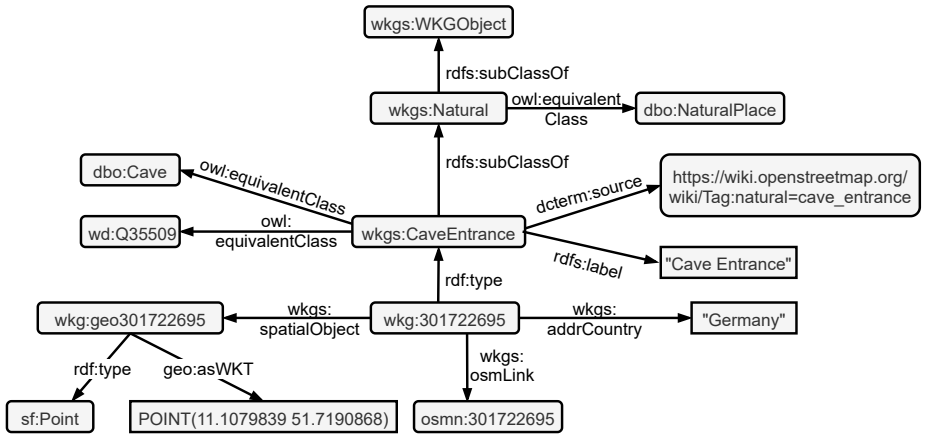


Figure 2: Example instantiation of the WorldKG ontology for a specific instance of `wkgs:CaveEntrance`.

```
wkg:1014675277 a wkgs:Restaurant;
   rdfs:label "Krishna" ;
   wkgs:addrCountry "DE" ;
   wkgs:addrHousenumber "53;54" ;
   wkgs:cuisine "indian" ;
   wkgs:dietVegetarian "yes";
   wkgs:openingHours "Mo-Su 17:00-23:00" ;
   wkgs:organic "only" ;
   wkgs:phone "+49 421 52279939" ;
   wkgs:spatialObject wkg:geo1014675227 ;
   wkgs:website "http://www.indisches-
                  bio-restaurant.de/" ;
   wkgs:wheelchair "no" ;
   wkgs:osmLink osmn:1014675277.

wkg:geo1014675227 a sf:Point;
   geo:asWKT "Point(8.7938916 53.073794)"
                    ^^geo:wktLiteral .
```

Listing 1: RDF Triples in the Turtle format for an example geographic entity of type `wkgs:Restaurant` in WorldKG.

5 WORLDKG CREATION PROCESS

In this section, we present our approach for creating WorldKG, consisting of the WorldKG ontology and geographic entities. First, we create the WorldKG ontology, which is then used to describe the geographic entities in WorldKG. The steps involved in the WorldKG creation process are depicted in Figure 3.

5.1 WorldKG Ontology Creation

The first part of the WorldKG creation process aims at creating the WorldKG ontology consisting of classes, properties, their relations, and links to the equivalent classes in Wikidata and DBpedia. This process consists of the following steps:

- *Scrape and filter key-value pairs:* First, we scrape the key-value pairs from the OSM map features which were introduced in Section 4. From these key-value pairs, we discard those that do not possess any class information. This concerns the key-value pairs categorized as *additional attributes*, *attributes* and *additional properties* in OSM map features.
- *Infer class hierarchy:* We use the keys to identify classes and key-value pairs to infer subclasses. If an individual value occurs with multiple keys, we manually specify a suitable subclass (e.g., for the key-value pairs `building=school` and `amenity=school`, we create classes *BuildingSchool* and *AmenitySchool*).
- *Convert property and class names:* To adhere to established OWL naming conventions [5], we represent WorldKG classes in upper camel-case format and properties in lower camel-case format.
- *Schema alignment with Wikidata and DBpedia.* We establish `owl:equivalentClass` relationships to Wikidata and DBpedia ontologies through the schema alignment process described in Section 4.

5.2 Knowledge Graph Creation

After the creation of the WorldKG ontology, we now utilize this ontology to represent OSM nodes as geographic entities in WorldKG. This process includes the following steps:

- *Filter nodes with at least one tag:* As input, we retrieve all OSM nodes from the most recent OSM dumps[10] using the Osmium Python library[11]. We filter out the nodes that do not contain any tags such as node:30519010[12]. These nodes are placeholders for ways and relations and are unlikely to be relevant for applications requiring node data.
- *Filter keys and values based on the WorldKG ontology:* Once we have collected the OSM nodes, we identify their classes and properties based on the WorldKG ontology and discard non-relevant tags and keys. From the OSM keys *lat* and *long*, we enrich the nodes with their geographic coordinates.
- *Create and validate triples:* Finally, we create RDF triples using the Python library *RDFLib* and provide links to the

[10] OSM dumps: https://download.geofabrik.de/
[11] Osmium python library: https://pypi.org/project/osmium/
[12] Filtered node: https://www.openstreetmap.org/node/30519010

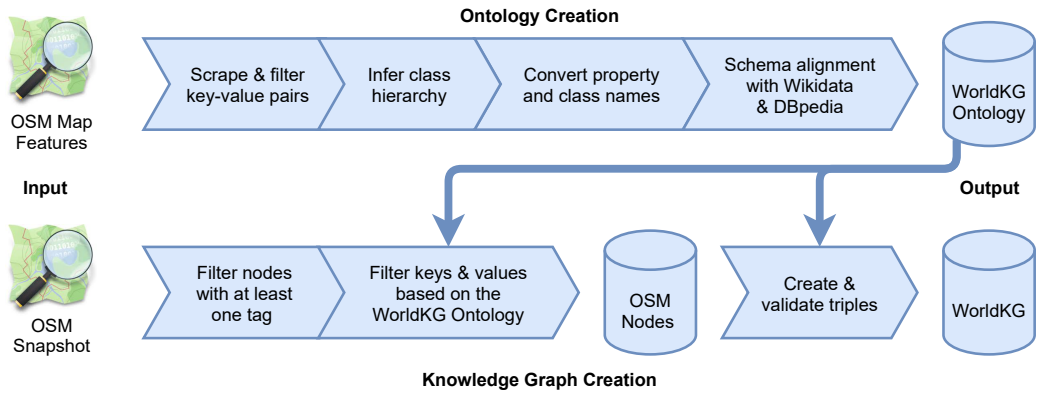


Figure 3: WorldKG ontology and knowledge graph creation process.

corresponding resources in Wikidata and DBpedia. Geographic objects are represented as `sf:Point` objects pointing to the coordinates as `geo:WKTLiteral` literals.

We provide an RDF dump of the geographic entities in WorldKG and its ontology. A SPARQL endpoint[13] using a Virtuoso triple store [13] is set up to query WorldKG.

6 WORLDKG CHARACTERISTICS & EVALUATION RESULTS

To illustrate the potential and quality of WorldKG, in this section, we present the statistics of the WorldKG and the evaluation results regarding the quality of the class alignment and type assertion.

6.1 WorldKG Statistics

As shown in Table 3, WorldKG contains more than 820 million triples associated with geographic data for 188 countries and seven continents. 33 top-level classes were inferred from OSM keys, whereas the subclasses refer to the specific classes extracted from key-value pairs, as discussed in Section 4.

Table 3: WorldKG knowledge graph statistics.

Quantity	Count
Total triples	828,550,751
Total entities	113,444,975
Top-level classes	33
Subclasses	1,143
Unique properties	1,820
Links to Wikidata classes	40
Links to DBpedia classes	21

6.2 Quality of the Class Alignment

As reported in [12], the NCA class alignment approach obtains matches with an average precision of 70% and 90% on the Wikidata and DBpedia knowledge graphs, respectively. As described in Section 4.2, we manually access the class alignments resulting from NCA and discard any wrong mappings to prevent the propagation of errors in WorldKG. By doing so, we obtain a class alignment precision of 100%. This manual verification procedure does not affect the recall values. This way, the recall corresponds to the original NCA recall of 63% and 81% on Wikidata and DBpedia knowledge graphs, respectively, reported in [12].

[13] WorldKG SPARQL endpoint: http://www.worldkg.org/sparql

6.3 Quality of the Type Assertion

In this section, we assess the quality of type assertion in WorldKG regarding the Wikidata and DBpedia classes. To this extent, we randomly select five classes from the DBpedia and the Wikidata ontologies mapped to WorldKG classes, respectively. For each of the resulting ten classes, we randomly select a sample of 100 WorldKG entities that are assigned to the respective class via `rdf:type` and `owl:equivalentClass`. Listing 2 shows the SPARQL query used for the generation of a sample dataset for the Wikidata class Q556186 labeled "mine". For each of the resulting 1000 entity-class pairs, we manually judge the correctness of the type assertion. That way, we can estimate the accuracy of the type assertion in WorldKG. The results are presented in Table 4.

```
SELECT ?id ?type ?osmid ?name
WHERE {
  ?id rdf:type ?type  .
  ?id rdfs:label ?name.
  ?id wkgs:osmLink ?osmid.
  ?type owl:equivalentClass wd:Q556186.
}
ORDER BY RAND() LIMIT 100
```

Listing 2: Query used to generate the sample set of 100 entities assigned to the Wikidata class Q556186 ("mine").

Tables 4a and 4b present the evaluation results of the type assertion of WorldKG entities to the Wikidata and DBpedia ontology classes, respectively. The *correct* and *wrong* columns indicate the

Table 4: Evaluation results of the WorldKG type assertion regarding Wikidata and DBpedia classes.

(a) Wikidata

WorldKG class	WorldKG entities	Wikidata class	Wikidata entities	Correct	Wrong	Non-verifiable	Accuracy
Tomb	12849	Q381885	3076	97	1	2	98.98%
Monument	44503	Q4989906	23320	91	0	9	100.00%
Mineshaft	8453	Q556186	677	95	2	3	97.94%
BicycleRental	40914	Q61663696	1757	96	0	4	100.00%
TourismHotel	204291	Q27686	11152	97	0	3	100.00%

(b) DBpedia

WorldKG class	WorldKG entities	DBpedia class	DBpedia entities	Correct	Wrong	Non-verifiable	Accuracy
ManMadeTower/ PowerTower	2769981	Tower	2533	97	0	3	100.00%
City	10465	City	22600	100	0	0	100.00%
Museum	46955	Museum	7422	94	2	4	97.92%
AmenitySchool	424236	School	31867	100	0	0	100.00%
CaveEntrance	39525	Cave	615	91	0	9	100.00%

number of correct or wrong KG classes assigned to individual entities, respectively. The *non-verifiable* column presents the number of cases in which we could not identify the correct class due to the lack of information about the entity on the web. For instance, an OSM node tagged with `historic=monument`, with no further information available, can not be verified to actually be a monument[14]. We exclude non-verifiable instances from our accuracy calculation. As we can observe, the precise tag-to-class mappings in WorldKG facilitate a very high accuracy (between 97.9% and 100%) of type assertion regarding both Wikidata and DBpedia classes. The few cases of incorrectly assigned classes result from wrongly annotated instances in OSM.

For all classes illustrated in Table 4a and Table 4b except of `wkgs:City`, the number of geographic entities in WorldKG is higher compared to Wikidata and DBpedia. Overall, as shown in Table 5, the number of geographic entities in WorldKG is two orders of magnitude higher than in Wikidata and DBpedia.

Overall, the high accuracy class alignment of the WorldKG pipeline builds the foundation for the integration of OSM information into the linked open data cloud. While OSM relies on the voluntarily contributed information, with no strict guarantees of correctness, WorldKG addresses this issue by only considering established tags defined in the OSM map feature list and therefore provides trustworthy high-quality information at scale.

Table 5: Number of geographic entities in WorldKG, Wikidata and DBpedia

Knowledge graph	Geographic entities
WorldKG	113,444,975
Wikidata	8,621,058
DBpedia	1,224,403

[14] An example non-verfiable node: https://www.openstreetmap.org/node/8752666922

7 AVAILABILITY, UTILITY & SUSTAINABILITY

In this section, we describe how the WorldKG website and the data and code repositories ensure the availability, utility, and sustainability of WorldKG.

7.1 Availability

The WorldKG website[15] is publicly available. This website provides a description of WorldKG and a SPARQL endpoint for querying the WorldKG knowledge graph. In addition, the WorldKG website provides pointers to code and data:

- *Code*: The code realizing the WorldKG creation process depicted in Figure 3 is available on GitHub[16] under the MIT License[17].
- *Data*: The WorldKG triples can be downloaded from a persistent URL[18] under the Open Data Commons Open Database License (ODbL)[19]. On the WorldKG website, we also made the manually created evaluation dataset of `owl:equivalentClass` mappings between the WorldKG ontology and the DBpedia/Wikidata ontologies available.

7.2 Utility

By following best practices in data publishing and open RDF W3C standard for modeling and interlinking the data, we envision WorldKG's establishment as part of the Linked Open Data Cloud. In detail, we ensure the utility of WorldKG via the following aspects:

[15] WorldKG Website: http://www.worldkg.org/
[16] GitHub Link: https://github.com/alishiba14/WorldKG-Knowledge-Graph
[17] License for code: https://opensource.org/licenses/MIT
[18] DOI for WorldKG data: https://doi.org/10.5281/zenodo.4953986
[19] License for dataset: https://opendatacommons.org/licenses/odbl/

```
PREFIX uom:
<http://www.opengis.net/def/uom/OGC/1.0/>

SELECT ?closeObject ?restaurant
 (bif:st_distance(?cWKT, ?fWKT, uom:metre)
  AS ?distance)
WHERE {
  ?poi rdfs:label "Brandenburger Tor".
  ?poi wkgs:spatialObject [
    geo:asWKT ?cWKT
  ] .
  ?closeObject rdf:type wkgs:Restaurant.
  ?closeObject rdfs:label ?restaurant.
  ?closeObject wkgs:spatialObject ?fGeom.
  ?fGeom geo:asWKT ?fWKT .
}
ORDER BY ASC(
  bif:st_distance(?cWKT, ?fWKT, uom:metre))
LIMIT 3
```

Listing 3: Example SPARQL query to retrieve the three closest restaurants to the Brandenburger Tor.

- *Documentation*: The WorldKG website provides a description of the data and the ontology. In addition, a selection of example SPARQL queries is given as an overview of potential usage scenarios and as a basis for creating new queries.
- *Data access*: WorldKG can be queried through its publicly available SPARQL endpoint that facilitates geographic queries with the option to download query results. Classes and resources of WorldKG can be looked up on the website, which also provides map visualizations pointing to the location of *wkgs:WKGObject* entities.
- *Provenance*: Version 1.0 of WorldKG was extracted using OSM dumps from June 6, 2021[20]. Among other metadata, this provenance information is provided as part of WorldKG using the VoID vocabulary [1]. Classes and instances in WorldKG are linked to Wikidata, DBpedia and OSM, where possible.

7.3 Sustainability

To keep WorldKG up-to-date with future releases of OSM that may further extend the coverage of real-world locations and account for potential transformations, we plan to publish new versions of WorldKG regularly. We further plan to add additional features in the upcoming versions of WorldKG, including enriched entity descriptions and extended coverage of real-world entities through data fusion with other sources.

8 EXAMPLE SCENARIO

This section demonstrates the usage of WorldKG for Point-of-Interest (POI) recommendation through an example scenario.

[20]The OSM dumps were downloaded from http://download.geofabrik.de/.

Table 6: Result of the example SPARQL query in Listing 3.

Restaurant	Distance
"Hopfingerbräu im Palais"	0.128322
"Restaurant Quarré"	0.243953
"Lorenz Adlon Esszimer"	0.247478

With the increased use of recreational and touristic applications, POI recommendation systems have gained increased attention [20, 22, 36]. Typically, the goal of a POI recommendation is to recommend a list of places to a user based on user-specific criteria, e.g., the user location and preferences. Knowledge graph, a machine-readable knowledge source supporting relational reasoning, can serve as a rich information source for POI recommendation [16]. However, cross-domain knowledge graphs often lack sufficient coverage of touristic POIs, such as restaurants. WorldKG fills this gap by providing means to retrieve POIs originated from OSM based on entity class labels.

Listing 3 exemplifies a SPARQL query that returns the three closest restaurants (`wkgs:Restaurant`) for a given location (the Brandenburger Tor in Berlin, Germany). This query makes use of GeoSPARQL functions (i.e., `bif:st_distance`) which are supported by the WorldKG SPARQL endpoint.

Table 6 shows the result of the example mentioned above after querying WorldKG, including the names of the restaurants and their distance to the Brandenburger Tor. Figure 4 shows a screenshot taken from the result page of the WorldKG SPARQL endpoint, which is the visualization of returned restaurants on a map. This example demonstrates how POI applications can immediately benefit from the provision of geographic information in WorldKG.

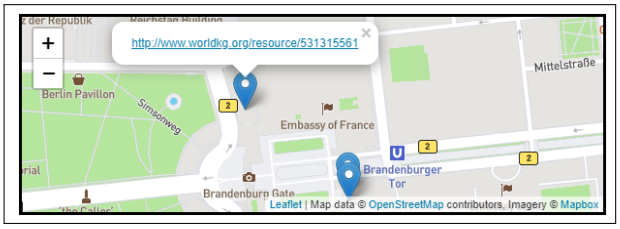


Figure 4: Visualization of the three restaurants closest to the Brandenburger Tor returned by the query in Listing 3.

9 RELATED WORK

In the following, we discuss existing KGs that include geographic entities, ontologies for geographic data and ontology alignment methods relevant for creating geographic KGs.

9.1 KGs containing Geographic Entities

There exist several dedicated geographic knowledge graphs as well as general-purpose knowledge graphs with geographic entities. Wang et al. [34] introduced GeoKG, a formalized geographic knowledge representation that complements the Attributive Language with Complements (ALC) description logic. Through case studies, the authors demonstrate that the GeoKG model can achieve

more accurate and complete geographic knowledge representations compared to YAGO.

DBpedia [2] was one of the first well-established, cross-domain knowledge graphs. To represent geospatial data, DBpedia provides latitude and longitude values for various geographic entities. Similarly, Wikidata [33] represents coordinates of geospatial entities. However, both DBpedia and Wikidata only cover a small fraction of OSM locations [32]. YAGO2geo [19] is an extension of the YAGO knowledge graph that includes geospatial and temporal relations. YAGO2geo is created using OSM and reference geospatial datasets such as Greek Administrative Geography (GAG) and Global Administrative Areas dataset (GADM). YAGO2geo mainly focuses on administrative regions and reuses an existing ontology from the GAG dataset. EventKG [14] is a knowledge graph that focuses on event-centric information and includes geographic entities relevant to historical events and their participants.

In general, the knowledge graphs mentioned above lack coverage of geospatial entities and focus on certain entity types. LinkedGeoData [3], on the other hand, converts OpenStreetMap data into an RDF knowledge graph. LinkedGeoData is based on a formal ontology created using tags and keys of OSM. it provides a simplified mapping between OSM data and classes and properties of other data sources. In contrast to WorldKG, LinkedGeoData uses a set of manually selected class mappings. Moreover, the latest available dumps of LinkedGeoData were released in 2015, and no links to Wikidata were provided.

9.2 Ontologies for Geographic data

There have been various approaches to build ontologies that cater to geographical data due to its unique structure. Sun et al. [31] have built a manual three-level ontology for geospatial data. Although the ontology they built can be reused, it is still incomplete and not assessed for quality. With OSM being one of the most prominent sources of open geographic information, there have been approaches to build ontologies catering to the OSM data structure: OSMOnto [8] describes OSM tags in an ontology that provides few links to existing ontologies such as schema.org. Similar to WorldKG, OSMOnto is represented as a class hierarchy extracted from OSM keys and values. Ballatore et al. [4] developed the OSM semantic network by crawling OSM Wiki pages. The network can be used to compute the similarity between the concepts and also for geospatial retrieval of entities, among others. In contrast to these works, WorldKG ontology is created in an automated way and covers a variety of geographic classes. Thus, it is flexible towards OSM updates and not limited in its coverage of geographic entities.

9.3 Ontology Alignment

Ontology alignment (also known as ontology matching) aims to establish correspondences between the elements of different ontologies. The efforts to interlink open semantic datasets and to benchmark ontology alignment approaches have been driven by the W3C Semantic Web Education and Outreach (SWEO) Linking Open Data community project [21] and the Ontology Alignment Evaluation Initiative (OAEI) [26] Ontology alignment is conducted at both the element-level and the structure-level [27]. The element-level alignment typically uses natural language descriptions of the ontology elements, such as labels and definitions. Element-level alignment adopts string similarity metrics such as edit distance. Structure-level alignment exploits the similarity of the neighboring ontology elements, including the taxonomy structure, as well as shared instances [24]. Element-level and structure-level alignments have also been adopted to align ontologies with relational data [10] and tabular data [37]. Jiménez-Ruiz et al. [18] divided the alignment task into independent, smaller sub-tasks, aiming to scale up to very large ontologies. Machine learning has been widely adopted for ontology alignment. In the GLUE architecture [11], semantic mappings are learned semi-automatically, while [25] proposed a matching system that integrates string-based and semantic similarity features. Recently, deep neural networks-based approaches have been used for ontology alignment and schema matching. Proposed architectures include convolutional neural networks [7], representation learning [30], and stacked autoencoders [35]. Until now, the lack of a well-defined ontology of OSM hindered the application of ontology alignment approaches to OSM data. WorldKG addresses this problem by providing alignments between OSM tags and knowledge graph classes. WorldKG builds upon the recently proposed Neural Class Alignment approach [12] that facilitates alignments between OSM tags and KG classes using a novel neural architecture.

10 CONCLUSION

In this paper, we presented WorldKG – a new geographic knowledge graph that provides semantic representations of geographic entities in the OpenStreetMap dataset. The released WorldKG knowledge graph contains over 828 million triples of over 100 million entities spread across 1176 classes. Through manual quality assessment performed on randomly selected sample data, we observe that WorldKG contains high accuracy data. We make the data dump available and provide a SPARQL endpoint for accessing WorldKG. By following best practices for semantic data publishing, we ensure the availability and usability of the data and are committed to maintaining regular updates for sustainability. We believe that WorldKG has the potential to aid many applications and future research that consume geographic data.

Acknowledgements. This work was partially funded by DFG, German Research Foundation under "WorldKG" (424985896).

REFERENCES

[1] Keith Alexander, Richard Cyganiak, Michael Hausenblas, and Jun Zhao. 2011. Describing Linked Datasets with the VoID Vocabulary. https://www.w3.org/TR/void/

[2] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary G. Ives. 2007. DBpedia: A Nucleus for a Web of Open Data. In *Proc. of the ISWC 2007 (Lecture Notes in Computer Science, Vol. 4825)*. Springer, 722–735.

[3] Sören Auer, Jens Lehmann, and Sebastian Hellmann. 2009. LinkedGeoData: Adding a spatial dimension to the web of data. In *Proc. of the ISWC 2009 (Lecture Notes in Computer Science, Vol. 5823)*. Springer, 731–746.

[4] Andrea Ballatore, Michela Bertolotto, and David C. Wilson. 2013. Geographic knowledge extraction and semantic similarity in OpenStreetMap. *Knowl. Inf. Syst.* 37, 1 (2013), 61–81.

[5] Sean Bechhofer, Frank van Harmelen, Jim Hendler, Ian Horrocks, Deborah L. McGuinness, Peter F. Patel-Schneider, and Lynn Andrea Stein. 2004. OWL Web Ontology Language. https://www.w3.org/TR/owl-ref/

[6] David Beckett, Tim Berners-Lee, Eric Prud'hommeaux, and Gavin Carothers. 2014. RDF 1.1 Turtle. https://www.w3.org/TR/turtle/

[7] Alexandre Bento, Amal Zouaq, and Michel Gagnon. 2020. Ontology Matching Using Convolutional Neural Networks. In *Proc. of the LREC 2020*. European Language Resources Association, 5648–5653.

[8] Mihai Codescu, Gregor Horsinka, Oliver Kutz, Till Mossakowski, and Rafaela Rau. 2011. Osmonto-an ontology of openstreetmap tags. *State of the map Europe (SOTM-EU)* 2011 (2011).

[9] Tarcísio Souza Costa, Simon Gottschalk, and Elena Demidova. 2020. Event-QA: A Dataset for Event-Centric Question Answering over Knowledge Graphs. In *Proc. of CIKM '20*. ACM, 3157–3164.

[10] Elena Demidova, Irina Oelze, and Wolfgang Nejdl. 2013. Aligning Freebase with the YAGO ontology. In *Proc. of the CIKM 2013*. ACM, 579–588.

[11] Anhai Doan, Jayant Madhavan, Pedro Domingos, and Alon Halevy. 2004. Ontology Matching: A Machine Learning Approach. In *Handbook on Ontologies*. Springer, 385–404.

[12] Alishiba Dsouza, Nicolas Tempelmeier, and Elena Demidova. 2021. Towards Neural Schema Alignment for OpenStreetMap and Knowledge Graphs. In *Proc. of the ISWC 2021 (Lecture Notes in Computer Science)*. Springer.

[13] Orri Erling and Ivan Mikhailov. 2009. RDF Support in the Virtuoso DBMS. In *Networked Knowledge - Networked Media - Integrating Knowledge Management, New Media Technologies and Semantic Systems*. Studies in Computational Intelligence, Vol. 221. 7–24.

[14] Simon Gottschalk and Elena Demidova. 2019. EventKG - the hub of event knowledge on the web - and biographical timeline generation. *Semantic Web* 10, 6 (2019), 1039–1070.

[15] Mordechai Haklay and Patrick Weber. 2008. OpenStreetMap: User-generated street maps. *IEEE Pervasive Comput.* 7, 4 (2008), 12–18.

[16] Lavdim Halilaj, Jürgen Lüttin, Susanne Rothermel, Santhosh Kumar Arumugam, and Ishan Dindorkar. 2021. Towards a knowledge graph-based approach for context-aware points-of-interest recommendations. In *Proc. of the SAC'21*. ACM, 1846–1854.

[17] Stephan Huber and Christoph Rust. 2016. Calculate travel time and distance with OpenStreetMap data using the Open Source Routing Machine (OSRM). *The Stata Journal* 16, 2 (2016), 416–423.

[18] Ernesto Jiménez-Ruiz, Asan Agibetov, Jiaoyan Chen, Matthias Samwald, and Valerie Cross. 2020. Dividing the Ontology Alignment Task with Semantic Embeddings and Logic-Based Modules. In *Proc. of ECAI 2020 (Frontiers in Artificial Intelligence and Applications, Vol. 325)*. IOS Press, 784–791.

[19] Nikolaos Karalis, Georgios M. Mandilaras, and Manolis Koubarakis. 2019. Extending the YAGO2 Knowledge Graph with Precise Geospatial Knowledge. In *Proc. of the ISWC 2019, Part II (Lecture Notes in Computer Science, Vol. 11779)*. Springer, 181–197.

[20] Wei Liu, Jing Wang, Arun Kumar Sangaiah, and Jian Yin. 2018. Dynamic metric embedding model for point-of-interest prediction. *Future Gener. Comput. Syst.* 83 (2018), 183–192.

[21] LOD. 2017. Linked Open Data. https://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

[22] Yingtao Luo, Qiang Liu, and Zhaocheng Liu. 2021. STAN: Spatio-Temporal Attention Network for Next Location Recommendation. In *Proc. of The Web Conference 2021 (WWW'21)*. ACM / IW3C2, 2177–2185.

[23] Pascal Neis. 2021. OSMstats. Retrieved 10-June-2021 from https://osmstats.neis-one.org/

[24] DuyHoa Ngo, Zohra Bellahsene, and Konstantin Todorov. 2013. Opening the Black Box of Ontology Matching. In *Proc. of the ESWC 2013 (Lecture Notes in Computer Science, Vol. 7882)*. Springer, 16–30.

[25] Ikechukwu Nkisi-Orji, Nirmalie Wiratunga, Stewart Massie, Kit-Ying Hui, and Rachel Heaven. 2018. Ontology Alignment Based on Word Embedding and Random Forest Classification. In *Proc. of the ECML PKDD 2018 (Lecture Notes in Computer Science, Vol. 11051)*. Springer, 557–572.

[26] OAEI. 2014. Ontology Alignment Evaluation Initiative. http://oaei.ontologymatching.org

[27] Lorena Otero-Cerdeira, Francisco Javier Rodríguez-Martínez, and Alma Gómez-Rodríguez. 2015. Ontology matching: A literature review. *Expert Syst. Appl.* 42, 2 (2015), 949–971.

[28] Dharmen Punjani et al. 2018. Template-based question answering over linked geospatial data. In *Proc. of the GIR@SIGSPATIAL 2018*. ACM, 7:1–7:10.

[29] Ross S. Purves, Paul Clough, Christopher B. Jones, Mark H. Hall, and Vanessa Murdock. 2018. Geographic Information Retrieval: Progress and Challenges in Spatial Search of Text. *Found. Trends Inf. Retr.* 12, 2-3 (2018), 164–318.

[30] Lirong Qiu, Jia Yu, Qiumei Pu, and Chuncheng Xiang. 2017. Knowledge entity learning and representation for ontology matching based on deep neural networks. *Clust. Comput.* 20, 2 (2017), 969–977.

[31] Kai Sun, Yunqiang Zhu, Peng Pan, Zhiwei Hou, Dongxu Wang, Weirong Li, and Jia Song. 2019. Geospatial data ontology: the semantic foundation of geospatial data integration and sharing. *Big Earth Data* 3, 3 (2019), 269–296.

[32] Nicolas Tempelmeier and Elena Demidova. 2021. Linking OpenStreetMap with knowledge graphs - Link discovery for schema-agnostic volunteered geographic information. *Future Gener. Comput. Syst.* 116 (2021), 349–364.

[33] Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A Free Collaborative Knowledgebase. *Commun. ACM* 57 (2014), 78–85.

[34] Shu Wang, Xueying Zhang, Peng Ye, Mi Du, Yanxu Lu, and Haonan Xue. 2019. Geographic Knowledge Graph (GeoKG): A Formalized Geographic Knowledge Representation. *ISPRS Int. J. Geo Inf.* 8, 4 (2019), 184.

[35] Chuncheng Xiang, Tingsong Jiang, Baobao Chang, and Zhifang Sui. 2015. ERSOM: A Structural Ontology Matching Approach Using Automatically Learned Entity Representation. In *Proc. of the EMNLP 2015*. The Association for Computational Linguistics, 2419–2429.

[36] Min Xie, Hongzhi Yin, Hao Wang, Fanjiang Xu, Weitong Chen, and Sen Wang. 2016. Learning Graph-based POI Embedding for Location-based Recommendation. In *Proc. of the CIKM 2016*. ACM, 15–24.

[37] Shuo Zhang and Krisztian Balog. 2020. Web Table Extraction, Retrieval, and Augmentation: A Survey. *ACM Trans. Intell. Syst. Technol.* 11, 2 (2020), 13:1–13:35.